



Zammit-Mangion, A., Cressie, N., Ganesan, A. L., O'Doherty, S., & Manning, A. J. (2015). Spatio-temporal bivariate statistical models for atmospheric trace-gas inversion. *Chemometrics and Intelligent Laboratory Systems*, 149, 227-241.  
<https://doi.org/10.1016/j.chemolab.2015.09.006>

Early version, also known as pre-print

Link to published version (if available):  
[10.1016/j.chemolab.2015.09.006](https://doi.org/10.1016/j.chemolab.2015.09.006)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

# Spatio-temporal bivariate statistical models for atmospheric trace-gas inversion

Andrew Zammit-Mangion<sup>a,\*</sup>, Noel Cressie<sup>a</sup>, Anita L. Ganesan<sup>b</sup>, Simon O' Doherty<sup>c</sup>, Alistair J. Manning<sup>d</sup>

<sup>a</sup>*National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics (SMAS), University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia*

<sup>b</sup>*School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK*

<sup>c</sup>*School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK*

<sup>d</sup>*Hadley Centre, Met Office, Fitzroy Road, Exeter, Devon EX1 3PB, UK*

---

## Abstract

Atmospheric trace-gas inversion refers to any technique used to predict spatial and temporal fluxes using mole-fraction measurements and atmospheric simulations obtained from deterministic computer models. Studies to date are most often of a data-assimilation flavour, which implicitly consider univariate statistical models with the flux as the variate of interest. This univariate approach typically assumes that the flux field is either a spatially correlated Gaussian process or a spatially uncorrelated non-Gaussian process with prior expectation fixed using flux inventories (e.g., NAEI or EDGAR in Europe). Here, we extend this approach in three ways. First, we develop a bivariate model for the mole-fraction field and the flux field. The bivariate approach allows optimal prediction of both the flux field and the mole-fraction field, and it leads to significant computational savings over the univariate approach. Second, we employ a lognormal spatial process for the flux field that captures both the lognormal characteristics of the flux field (when appropriate) and its spatial dependence. Third, we propose a new, geostatistical approach to incorporate the flux inventories in our updates, such that the posterior spatial distribution of the flux field is predominantly data-driven. The approach is illustrated on a case study of methane ( $\text{CH}_4$ ) emissions in the United Kingdom and Ireland.

**Keywords:** Conditional multivariate models, Methane emissions, Multivariate geostatistics, Hamiltonian Monte Carlo, Spatial statistics

---



---

\*Corresponding author

Email address: [azm@uow.edu.au](mailto:azm@uow.edu.au) (Andrew Zammit-Mangion)

## 1. Introduction

Atmospheric trace-gas inversion refers to any technique used to predict spatial and temporal fluxes of a gas from observations of mole fractions. Since mole-fraction measurements are affected by weather patterns that are time varying, there is no straightforward relationship between the observations and the fluxes. The relationship, termed a ‘source-receptor relationship’ (SRR), is typically obtained by simulating from a deterministic computer model, such as a Lagrangian Particle Dispersion Model (LPDM), which maps the fluxes to the mole fractions [1]. The SRR can be characterised through a bivariate function  $b_t(\mathbf{s}, \mathbf{u})$  where the spatial locations  $\mathbf{s}, \mathbf{u} \in \mathbb{R}^2$  are in a given domain of interest,  $D \subset \mathbb{R}^2$ , and  $t \in \{1, 2, \dots\}$  is a discrete-time index.

Figure 1, top-left panel, shows the UK and Ireland methane ( $\text{CH}_4$ ) total fluxes by grid cell, in units  $\text{g s}^{-1}$ , obtained from a combination of flux inventories. Here, the component inventories are principally the National Atmospheric Emissions Inventory (NAEI) [2] and the Emissions Database for Global Atmospheric Research version 4.2 (EDGAR) [3]; see [4] for details. Figure 1, top-right panel, shows an evaluation of  $b_t(\mathbf{s}, \mathbf{u})$ , in units  $\text{s ng}^{-1}$ , of methane fluxes, for  $\mathbf{s}$  located at a mole-fraction monitoring station at Angus, Scotland (TTA), where  $\mathbf{u}$  takes values on a discrete lattice, and where  $t = 1$  (01 January 2014 at midnight). Figure 1, bottom panel, shows the two-hourly averaged observations in parts per billion (ppb) at TTA, following background-removal (see Section 4.2 for details), for all of January 2014. In atmospheric trace-gas inversion, the aim is to recover a flux map, such as that in Figure 1, top-left panel, from time-series observations taken at various monitoring stations and from the collection of atmospheric simulations (one such is shown in Figure 1, top-right panel) that establish the SRR.

Atmospheric trace-gas inversion is an ill-posed problem. Consequently, “small uncertainties in the observational data correspond to much higher uncertainty in the emission[s]” [5]. However, in addition to observation and flux-field uncertainties, there is a third source of uncertainty arising from the use of an atmospheric model that does not perfectly match the true SRR due to physical parameterisations, solver discretisations, etc. Sometimes this third term is called the *discrepancy term*, and failure to acknowledge it can lead to over-confident predictions [6, 7]. Until very recently, this discrepancy was not considered separately from the observation error; see [4]. However, it is crucial to distinguish between the errors due to observations and those due to model misspecification, as these are likely to have different statistical properties.

A critical contribution of our work is to formalise the insight in [4] and treat the mole-fraction field as a second variable of interest. The consideration of modelling a mole-fraction field in

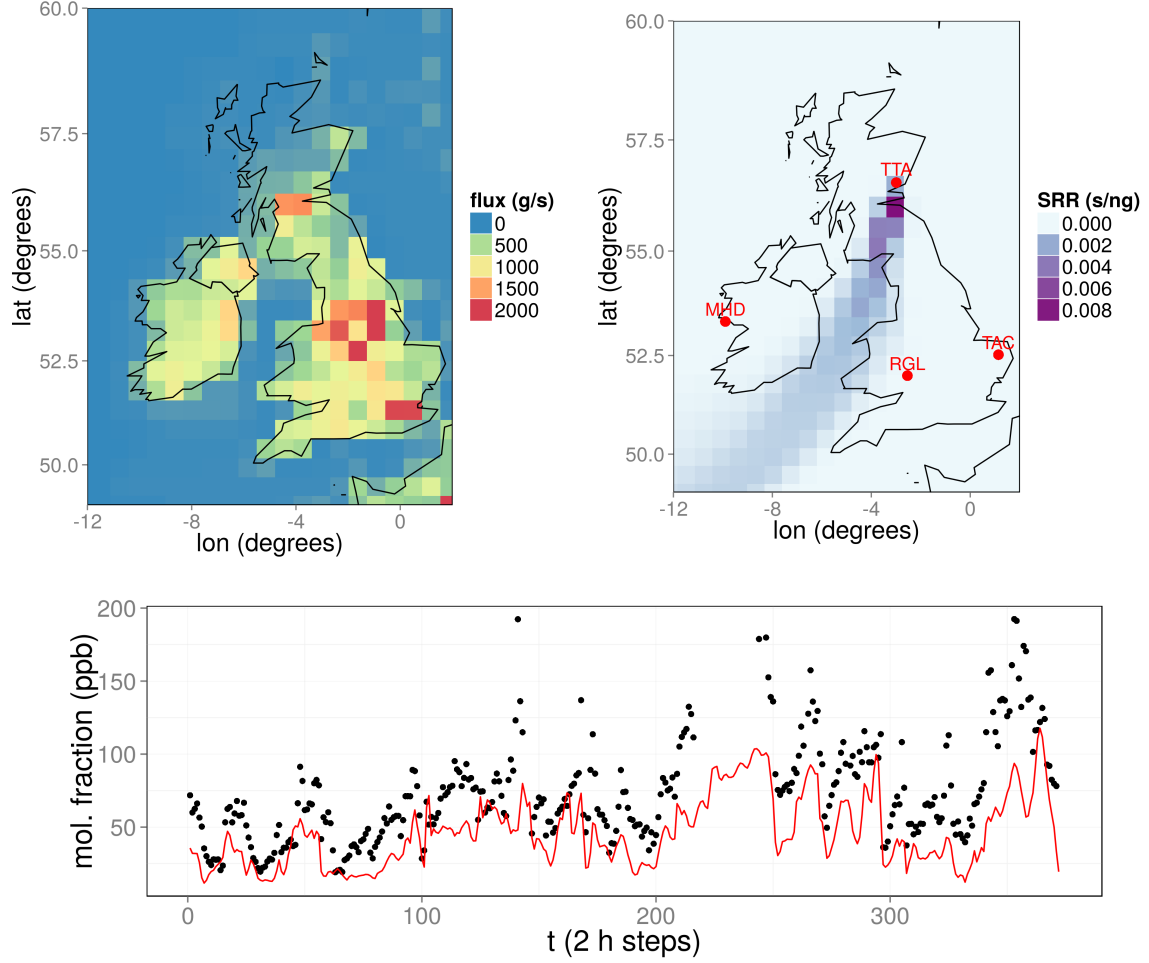


Figure 1: Top-left panel: Total flux by grid cell in  $\text{g s}^{-1}$  obtained by combining methane inventories (see [4] for details) for January 2014. Top-right panel: The source-receptor relationship on 01 January 2014 at 00:00,  $b_1(\mathbf{s}_{TTA}, \cdot)$ , obtained from the UK Met Office’s Lagrangian Particle Dispersion Model (LPDM), Numerical Atmospheric-dispersion Modelling Environment (NAME), where  $\mathbf{s}_{TTA}$  is the coordinate vector of the Angus measurement station (TTA), in Scotland. Also shown are the three other stations used in this study, Mace Head (MHD), Ridge Hill (RGL), and Tacolneston (TAC). Bottom panel: Measurements of methane mole fraction in parts per billion (ppb) at TAC for January 2014, following background-removal (black dots), together with a straightforward prediction (red line) using NAME and the methane flux inventory. Each time step corresponds to an interval of 2 h.

addition to the flux field through a discrepancy term results in a *bivariate* model (see [8] for a recent review on such models). The bivariate model brings two advantages to this problem of trace-gas inversion: The locations at which the mole-fraction field is modelled need not coincide with the data points. This in turn allows the predictive distribution of mole fraction at any unobserved locations to be found, and then averaged over any subset of the spatio-temporal domain, with relative ease (provided the SRR is available for these locations/domain). The other advantage is that the decoupling leads to computationally efficient methods in spatial statistics that can be used to scale up the inversion to large, remote-sensing datasets.

Another contribution of our work is to introduce the lognormal spatial process as a prior distribution for the flux field. This model acts as a bridge between the two sides of the dichotomous literature that either assumes spatial (possibly truncated) Gaussian-process priors (e.g., [9]) or spatially uncorrelated lognormal priors (e.g., [10]). A lognormal spatial process is attractive, as it is able to capture both (i) the nonnegativity and heavy tails in the distribution of the flux (valid for some trace gases such as methane) and (ii) the spatial correlation of the flux field. We show that expectations and covariances for both the flux field and the mole-fraction field can be obtained analytically if the flux field is defined as a lognormal process, by directly applying results from the univariate case [11, 12, p. 135].

The third contribution of our work is to propose a new way to carry out assimilation in atmospheric trace-gas inversion. Frequently, the prior expectation of the flux process is set from one or more inventories that are many times unreliable and that may have unquantifiable effects on a posterior assessment. Here, we only use *characteristics* of the inventories, namely the spatial length scales and the marginal variance, while setting the prior expectation to be spatially constant. In this way, the inventory fluxes are not used directly in the assimilation, and the spatial distribution of fluxes obtained from the posterior expectation will be predominantly data-driven. Our contribution addresses a concern in [13, Section 5.2] that suggests that such an approach is difficult when prior distributions are non Gaussian.

The article is structured as follows. In Section 2, we discuss the three contributions outlined above. In Section 3, we detail our approach to inferring the fields of interest using a combination of approximate inferential methods. The proposed framework is then assessed in Section 4, first in a study in one-dimensional space with simulated datasets, and then for emissions prediction in the UK and Ireland using the four measurement stations illustrated in Figure 1, top-right panel. Section 5 contains conclusions and an outline of future research directions.

## 2. Theory

The notation we use corresponds to that commonly found in the spatial-statistics literature (e.g., [14]). Here, stochastic processes are denoted using regular typeface, while bold typeface is used to denote vectors. The symbol  $Z$  denotes observations while  $Y$  denotes processes (fields). The notation  $Y(\cdot)$  is shorthand for ‘the process  $Y$  over the given spatial domain’ while  $Y(\mathbf{s})$  is the process  $Y$  evaluated at a specific location  $\mathbf{s}$ . The subscripts  $m$  and  $f$  are used for fields or domains associated with the mole-fraction field and the flux field, respectively, while the

subscript  $t$  denotes a discrete-time index. Therefore  $Y_f(\cdot)$  denotes the flux field, while  $Y_{m,t}(\cdot)$  denotes the spatio-temporal mole-fraction field at time-index  $t$ . Spatial domains are notated using the letter  $D$ ; the superscript  $O$  is used to denote observation domains and the superscript  $L$  is used to denote process (discretised onto a lattice) domains. The cardinality of a set is denoted using  $|\cdot|$ , while  $[A|B]$  denotes the conditional probability distribution of  $A$  given  $B$ .

For simplicity, as in [4, 10], we will assume throughout that the flux is a *spatial* process over a short time period; the development of models for *spatio-temporal* flux processes will be considered elsewhere. Recall that we use the subscript  $f$  to denote ‘flux,’ however because in this paper we constrain the fluxes to be positive *a priori*, we use the words ‘flux’ and ‘emissions’ interchangeably.

### 2.1. Bivariate modelling: Treating mole fraction as a second variate

In this section we discuss the univariate- and bivariate-modelling approaches to atmospheric trace-gas inversion, highlighting the key differences between the two. This discussion is used to motivate the advantages of the bivariate model, namely (i) interpretability, (ii) the possibility to compute the predictive distribution of mole fraction anywhere in, or averaged over any subset of, the domain of interest, and (iii) computational benefits arising from the decoupling.

Denote the mole-fraction observation at location  $\mathbf{s}$  and time  $t$  as  $Z_{m,t}(\mathbf{s})$ , and define  $D_{m,t}^O$  as the set of measurement locations at time  $t$ . Further, denote the collection of all observation locations as  $D_m^O \equiv \bigcup_{t \in \mathcal{T}} D_{m,t}^O$ , where  $\mathcal{T} \equiv \{1, \dots, T\}$  is the temporal domain and  $T > 1$  is the number of time steps. There are typically very few or no flux observations; flux prediction usually proceeds by using the following *univariate model* (or its discretised equivalent) for inversion [9, 15, 10, 16, 17]:

$$Z_{m,t}(\mathbf{s}) = \int_D b_t(\mathbf{s}, \mathbf{u}) Y_f(\mathbf{u}) d\mathbf{u} + e_{m,t}(\mathbf{s}); \quad \mathbf{s} \in D_{m,t}^O \subset D; \quad t \in \mathcal{T}, \quad (1)$$

where  $D$  is the spatial domain,  $e_{m,t}(\mathbf{s})$  is (Gaussian) observation error,  $b_t(\mathbf{s}, \mathbf{u})$  is the interaction function, or SRR, that returns quantities with dimensions  $[\text{time}][\text{mass}]^{-1}$ , and  $Y_f(\cdot)$  is the flux field that returns quantities with dimensions  $[\text{mass}][\text{time}]^{-1}[\text{area}]^{-1}$ . The field  $Y_f(\cdot)$  is a stochastic process that, as we discuss below in Section 2.2, has statistical properties that reflect those of the underlying process (e.g., positivity). Note that  $\{Y_f(\mathbf{u}) : \mathbf{u} \in D\}$  is the flux density, and it is different from the flux by grid cell  $\{A(\mathbf{u})Y_f(\mathbf{u}) : \mathbf{u} \in D^L\}$  (where  $A(\mathbf{u})$  is the grid cell area at  $\mathbf{u}$  and  $D^L$  is some discrete subset of  $D$ ) that returns quantities with dimensions

$[\text{mass}][\text{time}]^{-1}$ . Note also that there is no mention of the complete mole-fraction field in (1), only observations of the mole-fraction field at specific locations.

In our *bivariate model*, the object of interest shifts from the *observations* of mole fraction to the mole-fraction *field* itself, which we model as

$$Y_{m,t}(\mathbf{s}) = \int_D b_t(\mathbf{s}, \mathbf{u}) Y_f(\mathbf{u}) d\mathbf{u} + \zeta_t(\mathbf{s}); \quad \mathbf{s} \in D_m^L; \quad t \in \mathcal{T}, \quad (2)$$

where  $Y_{m,t}(\cdot)$  is the mole-fraction field and  $D_m^L$  is a user-defined set of discrete locations in  $D$  at which  $Y_{m,t}$  is evaluated. The additive component,  $\zeta_t(\mathbf{s})$ , captures the spatio-temporal discrepancy (due to mole-fraction background) and variation in the true mole-fraction field that cannot be explained by the flux field through the SRR. Note that it is different from  $e_{m,t}(\mathbf{s})$  in (1), which represents all the unexplained signal, since it explicitly accounts for *model* misspecification. Finally, the observations available for the inversion are not the field itself, but a noisy, incomplete version of it:

$$Z_{m,t}(\mathbf{s}) = Y_{m,t}(\mathbf{s}) + \varepsilon_{m,t}(\mathbf{s}); \quad \mathbf{s} \in D_{m,t}^O, \quad (3)$$

where  $\varepsilon_{m,t}(\cdot)$  is a measurement-error process.

There are three important differences between (1) and (2). First, (2) is defined on  $D_m^L$ , while (1) is only defined for  $D_{m,t}^O \subset D$ ; that is, the locations at which mole fraction is inferred need not (but could) coincide with the observation locations. Second, due to the decoupling, the locations  $\mathbf{s} \in D_m^L$  at which the SRR is evaluated are not dependent on the observation locations; the case of  $|D_m^L| \ll |D_m^O|$  occurs for regional studies that use computationally-intensive LPDMs to simulate the SRR. Third, the mole-fraction observations do not appear in (2). This is computationally beneficial as it is usually reasonable to assume that observations are conditionally independent given the mole-fraction field, but not necessarily given only the flux field. This conditional independence implies that the conditional covariance matrix of  $\{Z_{m,t}(\mathbf{s})\}$  given the mole-fraction field is diagonal, resulting in simplified matrix computations.

Consider a discretisation of  $D$  for the flux field, denoted  $D_f^L$  (that need not be identical to  $D_m^L$ ). Define

$$\begin{aligned}
\mathbf{Z}_{m,t} &\equiv (Z_{m,t}(\mathbf{s}) : \mathbf{s} \in D_{m,t}^O)', & \mathbf{e}_{m,t} &\equiv (e_{m,t}(\mathbf{s}) : \mathbf{s} \in D_{m,t}^O)', \\
\mathbf{Z}_m &\equiv (\mathbf{Z}'_{m,t} : t \in \mathcal{T})', & \mathbf{e}_m &\equiv (\mathbf{e}'_{m,t} : t \in \mathcal{T})', \\
\mathbf{Y}_{m,t} &\equiv (Y_{m,t}(\mathbf{s}) : \mathbf{s} \in D_m^L)', & \boldsymbol{\varepsilon}_{m,t} &\equiv (\varepsilon_{m,t}(\mathbf{s}) : \mathbf{s} \in D_{m,t}^O)', \\
\mathbf{Y}_m &\equiv (\mathbf{Y}'_{m,t} : t \in \mathcal{T})', & \boldsymbol{\varepsilon}_m &\equiv (\boldsymbol{\varepsilon}'_{m,t} : t \in \mathcal{T})', \\
\mathbf{Y}_f &\equiv (Y_f(\mathbf{s}) : \mathbf{s} \in D_f^L)', & \boldsymbol{\zeta}_t &\equiv (\zeta_t(\mathbf{s}) : \mathbf{s} \in D_m^L)', \\
& & \boldsymbol{\zeta} &\equiv (\boldsymbol{\zeta}'_t : t \in \mathcal{T})',
\end{aligned}$$

where the ‘prime’ symbol is used to denote the transpose operation. Then, in the univariate model (1), the relationship between the mole-fraction *observations* and the flux field is approximated by

$$\mathbf{Z}_{m,t} \simeq \mathbf{B}_{U,t} \mathbf{Y}_f + \mathbf{e}_{m,t}; \quad t \in \mathcal{T}, \quad (4)$$

where  $\mathbf{B}_{U,t} \equiv (A(\mathbf{u})b_t(\mathbf{s}, \mathbf{u}) : \mathbf{s} \in D_{m,t}^O; \mathbf{u} \in D_f^L)$  is a  $|D_{m,t}^O| \times |D_f^L|$  matrix that approximates the integral in (1), and  $\{A(\mathbf{u}) : \mathbf{u} \in D_f^L\}$  are integration weights with dimension, [area]. In the bivariate model, the relationship between the mole-fraction *field* and the flux field, (2), is approximated by

$$\mathbf{Y}_{m,t} \simeq \mathbf{B}_{B,t} \mathbf{Y}_f + \boldsymbol{\zeta}_t; \quad t \in \mathcal{T}, \quad (5)$$

where

$$\mathbf{B}_{B,t} \equiv (A(\mathbf{u})b_t(\mathbf{s}, \mathbf{u}) : \mathbf{s} \in D_m^L; \mathbf{u} \in D_f^L) \quad (6)$$

is a  $|D_m^L| \times |D_f^L|$  matrix that approximates the integral in (2). The relationship between the mole-fraction observations and the mole-fraction field is

$$\mathbf{Z}_{m,t} = \mathbf{C}_t \mathbf{Y}_{m,t} + \boldsymbol{\varepsilon}_{m,t}; \quad t \in \mathcal{T}, \quad (7)$$

where  $\mathbf{C}_t$  is a  $|D_{m,t}^O| \times |D_m^L|$  incidence matrix that indicates where the observation locations  $D_{m,t}^O$  are in  $D_m^L$ .

The first advantage of the bivariate approach is the ability to compute the predictive distribution of the mole-fraction field at locations, or over space-time domains, where the field is not directly observed (note that the SRR still needs to be available for these locations or domains). In the univariate approach, (4), we are able to compute (through a standard application of



Bayes' rule), the probability distribution of the flux field given the mole-fraction observations, that is,  $[\mathbf{Y}_f | \mathbf{Z}_m]$ . Following this, we could try to obtain the predictive distribution of the mole fractions at time  $t$  through  $[\mathbf{B}_{U,t} \mathbf{Y}_f | \mathbf{Z}_m]$ , but this does not take into account the discrepancy  $\zeta_t(\mathbf{s})$ . In the bivariate case, with (5) and (7) we can compute the *joint* posterior distribution  $[\mathbf{Y}_f, \mathbf{Y}_m | \mathbf{Z}_m]$  at locations  $D_m^L \times D_f^L$ , from which we can in principle obtain both the posterior distribution of the flux field  $[\mathbf{Y}_f | \mathbf{Z}_m]$  *and* that of the mole-fraction field  $[\mathbf{Y}_m | \mathbf{Z}_m]$ . We may also easily find the posterior distribution of any affine transformation of the random vector  $\mathbf{Y}_m$ , and hence it is straightforward to carry out inferences over a spatial or temporal aggregation of the mole-fraction field if desired.

The second advantage of the bivariate approach is computational. In many applications it is reasonable to assume that the observation error  $\boldsymbol{\varepsilon}_m$  is uncorrelated. However, this assumption does not hold for  $\mathbf{e}_m$  (in the univariate case), since  $e_{m,t}(\mathbf{s})$  should incorporate spatio-temporal correlations arising indirectly from the discrepancy term [10]. Since the observations  $\mathbf{Z}_m$  are (in practice) spatio-temporally irregular, there is no straightforward way, without modifying the underlying model (e.g., through covariance tapering [18]), to define  $\text{cov}(\mathbf{e}_{m,t})$  such that it is sparse or such that it can be decomposed into simpler components. On the other hand,  $Y_{m,t}(\cdot)$  (equivalently  $\zeta_t(\cdot)$ ) can be discretised in such a way that covariance matrices associated with  $\zeta_t(\mathbf{s})$  can be easily stored and inverted. For example, in Section 3.2, we use a regular space-time grid for the discretisation, and we exploit the properties of the Kronecker product to simplify the computation as in [4].

## 2.2. The lognormal bivariate model and its properties

To date, most studies have assumed either that  $Y_f(\cdot)$  is a Gaussian process or that  $Y_f(\cdot)$  is spatially uncorrelated and non-Gaussian. Non-Gaussianity is generally important to model since the spatial distribution of fluxes tends to exhibit a heavy right tail and, for some trace gases such as methane, it is reasonable to further assume that it has only positive support. However, it is easy to see from flux inventories that often the spatially-uncorrelated assumption is unrealistic. In this section we present a new model, namely a lognormal spatial process, that can cater for both non-Gaussianity and for spatial correlations that appear in the flux field. We then embed this in the bivariate model and discuss the model's resulting first-order and second-order properties.

Assume that  $Y_f(\cdot)$  is a spatial process such that  $\tilde{Y}_f(\cdot) \equiv \ln Y_f(\cdot)$  is a Gaussian process [19].

Then  $Y_f(\cdot)$  is termed a *lognormal spatial process*; clearly,  $P(Y_f(\mathbf{s}) \leq 0) = 0$ , for  $\mathbf{s} \in D$ , so that  $Y_f(\cdot)$  is almost surely positive everywhere. Assume further that  $E(\tilde{Y}_f(\mathbf{s}) | \boldsymbol{\vartheta}) \equiv \tilde{\mu}_f(\mathbf{s} | \boldsymbol{\vartheta})$  and  $\text{cov}(\tilde{Y}_f(\mathbf{s}), \tilde{Y}_f(\mathbf{u}) | \boldsymbol{\vartheta}) \equiv \tilde{C}_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta})$ , where  $\boldsymbol{\vartheta}$  is a vector of unknown parameters. Then from [20],

$$\mu_f(\mathbf{s} | \boldsymbol{\vartheta}) \equiv E(Y_f(\mathbf{s}) | \boldsymbol{\vartheta}) \equiv \exp(\tilde{\mu}_f(\mathbf{s} | \boldsymbol{\vartheta}) + (1/2)\tilde{C}_{ff}(\mathbf{s}, \mathbf{s} | \boldsymbol{\vartheta})); \quad \mathbf{s} \in D, \quad (8)$$

$$C_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta}) \equiv \text{cov}(Y_f(\mathbf{s}), Y_f(\mathbf{u}) | \boldsymbol{\vartheta}) \equiv \mu_f(\mathbf{s} | \boldsymbol{\vartheta})\mu_f(\mathbf{u} | \boldsymbol{\vartheta})[\exp(\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta})) - 1]; \quad \mathbf{s}, \mathbf{u} \in D. \quad (9)$$

That is, the mean and covariances of the lognormal spatial process are adjusted versions of the exponentiated mean and covariances of the Gaussian spatial process, and thus they can be easily computed.

We only need to specify two components in (8) and (9): the prior mean and the prior covariance functions in the ‘log space,’ that is,  $\tilde{\mu}_f(\mathbf{s} | \boldsymbol{\vartheta})$  and  $\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta})$ ;  $\mathbf{s}, \mathbf{u} \in D$ . The way in which we do so does not differ substantially from when Gaussian spatial processes are used to model the prior and, in this regard, research to date falls into two broad categories. First, in works that employ a data-assimilation framework,  $\tilde{\mu}_f(\mathbf{s} | \boldsymbol{\vartheta})$  is typically set from values in a flux inventory, whilst  $\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta})$ ;  $\mathbf{s}, \mathbf{u} \in D$ , is chosen through expert judgment (e.g., [21]). Second, in works that employ a geostatistical framework,  $\tilde{\mu}_f(\mathbf{s} | \boldsymbol{\vartheta})$  is defined as a linear combination of regressors, whilst  $\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} | \boldsymbol{\vartheta})$ ;  $\mathbf{s}, \mathbf{u} \in D$ , is a standard univariate covariance function (e.g., the stationary and isotropic exponential covariance function [9, 22]). In Section 2.3, we offer an enhanced approach to eliciting the first-order and second-order moments of  $Y_f(\cdot)$ .

Given (2), we are able to obtain closed-form expressions for the expectation and covariance function of the mole-fraction field at time  $t$ . These are,

$$\mu_{m,t}(\mathbf{s} | \boldsymbol{\vartheta}) \equiv E(Y_{m,t}(\mathbf{s}) | \boldsymbol{\vartheta}) = \int_D b_t(\mathbf{s}, \mathbf{v})\mu_f(\mathbf{v} | \boldsymbol{\vartheta})d\mathbf{v} + E(\zeta_t(\mathbf{s})); \quad \mathbf{s} \in D, \quad (10)$$

where we deliberately allow the possibility that  $E(\zeta_t(\mathbf{s})) \neq 0$ ; and

$$\begin{aligned} C_{mm,t}(\mathbf{s}, \mathbf{u} | \boldsymbol{\theta}, \boldsymbol{\vartheta}) &\equiv \text{cov}(Y_{m,t}(\mathbf{s}), Y_{m,t}(\mathbf{u}) | \boldsymbol{\theta}, \boldsymbol{\vartheta}) \\ &= C_{\zeta\zeta,t}(\mathbf{s}, \mathbf{u} | \boldsymbol{\theta}) + \int_D \int_D b_t(\mathbf{s}, \mathbf{v})C_{ff}(\mathbf{v}, \mathbf{w} | \boldsymbol{\vartheta})b_t(\mathbf{u}, \mathbf{w})d\mathbf{v}d\mathbf{w}; \quad \mathbf{s}, \mathbf{u} \in D, \end{aligned} \quad (11)$$

where  $C_{\zeta\zeta,t}(\mathbf{s}, \mathbf{u} | \boldsymbol{\theta}) \equiv \text{cov}(\zeta_t(\mathbf{s}), \zeta_t(\mathbf{u}) | \boldsymbol{\theta})$ , and  $\boldsymbol{\theta}$  is a vector of unknown parameters appearing in the covariance function of the discrepancy term. Again, from (2), the cross-covariance

functions between the mole-fraction field at time  $t$  and the flux field are,

$$C_{fm,t}(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\vartheta}) \equiv \text{cov}(Y_f(\mathbf{s}), Y_{m,t}(\mathbf{u}) \mid \boldsymbol{\vartheta}) = \int_D C_{ff}(\mathbf{s}, \mathbf{w} \mid \boldsymbol{\vartheta}) b_t(\mathbf{u}, \mathbf{w}) d\mathbf{w}; \quad \mathbf{s}, \mathbf{u} \in D, \quad (12)$$

$$C_{mf,t}(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\vartheta}) \equiv \text{cov}(Y_{m,t}(\mathbf{s}), Y_f(\mathbf{u}) \mid \boldsymbol{\vartheta}) = \int_D C_{ff}(\mathbf{w}, \mathbf{u} \mid \boldsymbol{\vartheta}) b_t(\mathbf{s}, \mathbf{w}) d\mathbf{w}; \quad \mathbf{s}, \mathbf{u} \in D. \quad (13)$$

Equations (10)–(13) are identical to those given in [23] with  $\mu_f$  and  $C_{ff}$  given by (8) and (9), respectively. Now we can write down the expectation and covariance of the joint model at time  $t$  as

$$\mathbf{Y}_t(\cdot) \sim \text{Dist} \left( \begin{pmatrix} \mu_f(\cdot) \\ \mu_{m,t}(\cdot) \end{pmatrix}, \begin{pmatrix} C_{ff}(\cdot, \cdot) & C_{fm,t}(\cdot, \cdot) \\ C_{mf,t}(\cdot, \cdot) & C_{mm,t}(\cdot, \cdot) \end{pmatrix} \right), \quad (14)$$

where  $\mathbf{Y}_t(\cdot) \equiv (Y_f(\cdot), Y_{m,t}(\cdot))'$ ,  $\text{Dist}(\cdot, \cdot)$  is a distribution that here is not Gaussian but which features the first two moments, and the terms in (14) are given by (10)–(13). We stress that, unlike in a bivariate Gaussian process, here the joint model  $[\mathbf{Y}_t(\cdot)]$  given by (14) is not fully specified by the mean and covariance functions. However, closed-form equations (8) and (9), for the mean and covariance functions, are useful for developing computationally efficient approximate inference methods, such as those described in Section 3.

We have derived above the expectation and covariance of  $\mathbf{Y}_t(\cdot)$ , that is, of the bivariate process at a single time point  $t$ . In practice, we need to find the expectation and covariance of

$$\mathbf{Y}(\cdot) \equiv (Y_f(\cdot), (Y_{m,t}(\cdot) : t \in \mathcal{T}))'. \quad (15)$$

These can be obtained using the same ideas used to construct (10)–(13) and the mole-fraction covariances  $\text{cov}(Y_{m,t}(\cdot), Y_{m,t'}(\cdot)); t, t' \in \mathcal{T}$ .

### 2.3. Inventories for estimating the spatial lognormal characteristics of the flux field

One would ideally estimate from the data all unknown parameters appearing in the model. However, this can be difficult for parameters affecting fields that are not directly observed (in our case  $Y_f(\cdot)$ ) and in ill-posed problems such as that considered here. In other fields of study, this problem has been rectified by estimating the parameters from deterministic models. In [24], for example, parameters appearing in a dynamic stochastic model of aerosol optical depth were estimated from a chemical-tracer numerical model. In [25], the varying spatial-length scales of precipitation patterns were estimated from a regional climate model. In the present study, we

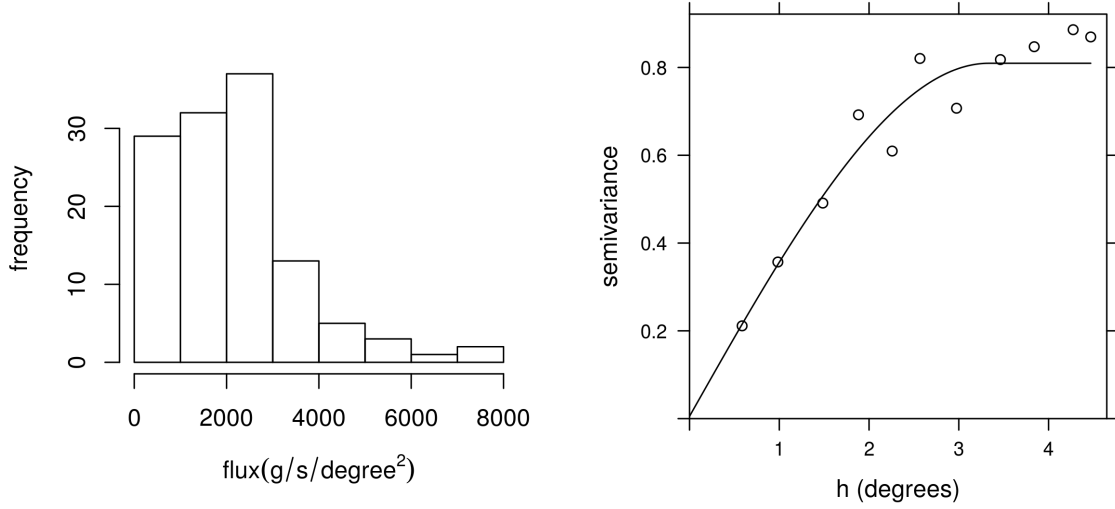


Figure 2: Left panel: Histogram of total methane flux by grid cell in the UK and Ireland (see Figure 1, top-left panel). Right panel: The empirical (open circles) and fitted (solid line) isotropic semivariogram of the methane flux as a function of lag distance in degrees (latitude and longitude).

propose estimating the spatial-length scale and the marginal variance of the process  $Y_f(\cdot)$  from flux inventories. Importantly, unlike several works in atmospheric trace-gas inversion, we do not fix the prior expectation of the flux process to be equal to values given in an inventory, since it is often found that inventories can be biased due to errors in activity data or emissions factors. Instead, we set  $\tilde{\mu}_f(\cdot) = c$ , where  $c \in \mathbb{R}$  is a spatially-invariant constant evaluated below. We now show the procedure for obtaining  $c$  and the parameters in  $\tilde{C}_{ff}(\cdot, \cdot)$  from the inventory shown in Figure 1, top-left panel.

The flux field is generally spatio-temporal and not isotropic, but here we only seek to capture the dominant spatial dependence using a simple spatial isotropic model; the spatio-temporal discrepancy term  $\zeta_i(\mathbf{s})$  accounts for any remaining variability. First, we divide the total flux by grid cell in our inventory (shown in Figure 1, top-left panel) by the grid cell area in order to obtain flux densities, which we collect into a vector  $\mathbf{Z}_{inv}$ . Then, we let  $\mathbf{Y}_{inv} \equiv \ln \mathbf{Z}_{inv}$ , elementwise. Note that  $\mathbf{Y}_{inv}$  can be considered dimensionless since, for any constant  $k$ ,  $\ln(kZ_{inv}^i) = \ln k + \ln Z_{inv}^i$ , where  $Z_{inv}^i$  is the  $i$ th element of  $\mathbf{Z}_{inv}$ . A histogram of the elements in  $\mathbf{Z}_{inv}$  is given in Figure 2, left panel, while the empirical isotropic semivariogram based on  $\mathbf{Y}_{inv}$  is shown in Figure 2, right panel.

To obtain  $\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\vartheta})$ , we first fit a spherical semivariogram,  $\gamma^o(h \mid \boldsymbol{\vartheta})$ , to the empirical semivariogram shown in Figure 2, where  $h = \|\mathbf{u} - \mathbf{s}\|$ , and  $\boldsymbol{\vartheta}$  is the vector of parameters

appearing in the specification of the semivariogram. Parameter estimates  $\hat{\boldsymbol{\vartheta}}$  were obtained using weighted least squares with `fit.variogram` in the package `gstat` with *R Software* [26, 27]. We fitted three semivariogram models: the spherical, the exponential, and the Gaussian semivariogram. From these three, the spherical model was the one that gave the best fit in terms of smallest sum-of-squared errors and thus is the one we proceeded with in our analysis. For the spherical semivariogram,  $\boldsymbol{\vartheta} \equiv (\sigma_1^2, \sigma_2^2, R)'$ ,  $\sigma_1^2$  is the dimensionless variance of a fine-scale variation component (nugget),  $\sigma_2^2$  is the dimensionless variance of a spatially correlated component (partial sill),  $R$  is a range parameter in degrees (latitude and longitude), and

$$\gamma^o(h \mid \boldsymbol{\vartheta}) \equiv \begin{cases} 0, & h = 0; \\ \sigma_1^2 + \sigma_2^2((3h)/(2R) - (1/2)(h/R)^3), & 0 < h < R; \\ \sigma_1^2 + \sigma_2^2, & h \geq R. \end{cases} \quad (16)$$

The spherical semivariogram has a finite sill,  $\sigma_1^2 + \sigma_2^2$ , and hence there is a stationary covariance function [12, Section 2.3.2],

$$\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\vartheta}) \equiv \sigma_1^2 + \sigma_2^2 - \gamma^o(\|\mathbf{u} - \mathbf{s}\| \mid \boldsymbol{\vartheta}). \quad (17)$$

The parameters corresponding to the fitted semivariogram, shown in Figure 2, right panel, are  $\hat{\boldsymbol{\vartheta}} = (0.0053, 0.80, 3.3^\circ)'$ . The fitted values were used in (17) to obtain the covariance function,  $\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} \mid \hat{\boldsymbol{\vartheta}})$ , which is used to describe the spatial dependence in the log-flux field. We complete the prior specification of the flux field by recalling that  $\tilde{\mu}_f(\cdot) = c$ , and setting  $c = \sum_{\mathbf{s} \in D_f^L} Y_{inv}(\mathbf{s}) / |D_f^L| = 7.35$ , which is a spatially invariant, dimensionless constant. We verified the adequacy of the fitted lognormal process through a leave-one-out cross-validation study and analysis of the prediction residuals in the log space.

### 3. Methods

Atmospheric trace-gas-inversion problems are computationally demanding and the bivariate model does not fully alleviate this. First, there are four domains to consider,  $\mathcal{T}$ ,  $D_m^O$ ,  $D_m^L$ , and  $D_f^L$ , the sizes of which can affect the computational efficiency and hence the strategy we employ. Second, even after estimating  $\boldsymbol{\vartheta}$  using the inventories, there are still unknown parameters in the discrepancy that need to be estimated; we denote those parameters as  $\boldsymbol{\theta}$ . For small problems, a full Markov chain Monte Carlo (MCMC) procedure to sample from the posterior distribution,

$[\mathbf{Y}_f, \mathbf{Y}_m, \boldsymbol{\theta} \mid \mathbf{Z}_m, \hat{\boldsymbol{\vartheta}}]$ , may be employed [4] and, when  $|\mathcal{T}|$  grows, exact or approximate algorithms that process the data one-temporal-interval-at-a-time may be used [28, 29].

Hierarchical models where prior distributions are placed on the unknown parameters  $\boldsymbol{\theta}$ , are known as *Bayesian* hierarchical models (BHMs). As both  $|D_m^L|$  and  $|D_f^L|$  increase, MCMC schemes may become prohibitive due to the time taken to obtain a single sample from the posterior distribution and due to poor mixing resulting from increased correlations between  $\boldsymbol{\theta}$  and  $(\mathbf{Y}'_f, \mathbf{Y}'_m)'$ . To deal with this, one can implement *empirical* hierarchical models (EHMs, see [14, pp. 20–21]) in which the parameters  $\boldsymbol{\theta}$  are treated as fixed unknowns that are estimated using standard maximum-likelihood techniques. MCMC is then used to sample from the empirical predictive distribution,  $[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ , where the estimate  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$  is substituted into the posterior distribution in place of  $\boldsymbol{\theta}$ . EHMs do not account for ‘parameter uncertainty’ in the Bayesian sense, and prediction intervals obtained for  $[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$  tend to be narrower than those obtained using BHMs [10, 30]. Nonetheless, EHMs have been shown to give reasonable results at a fraction of the computational cost of BHMs [31]. We discuss how adjusted intervals may be obtained for EHMs in Section 5.

### 3.1. Estimating the remaining parameters

Since the flux and mole-fraction fields are not directly observed, the likelihood of  $\boldsymbol{\theta}$  is not available in closed form. Maximum-likelihood estimation in this context can be treated as a ‘missing-data’ problem. Missing-data problems appear in all statistical contexts, and they are often solved using iterative methods. One of these methods is the expectation-maximisation (EM) algorithm, given in the celebrated article of [32]. Since its appearance, the EM algorithm has been widely employed for parameter estimation in empirical hierarchical spatio-temporal models (e.g., [33]), where the latent process is treated as missing data.

The EM algorithm for estimating  $\boldsymbol{\theta}$  is relatively straightforward to implement. Define the complete-data likelihood,  $L_c(\boldsymbol{\theta}) \equiv [\mathbf{Y}_f, \mathbf{Y}_m, \mathbf{Z}_m \mid \boldsymbol{\theta}, \hat{\boldsymbol{\vartheta}}]$ . Thus,  $L_c(\boldsymbol{\theta})$  inherits its randomness from  $\mathbf{Y}_f, \mathbf{Y}_m$ , and  $\mathbf{Z}_m$ . The EM algorithm is defined iteratively through: (i) the E-step, in which the function

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) \equiv \mathbb{E}(\ln L_c(\boldsymbol{\theta}) \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}),$$

is found for some ‘current’ estimate  $\boldsymbol{\theta}^{(i)}$ , and (ii) the M-step, which computes

$$\boldsymbol{\theta}^{(i+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}). \quad (18)$$

The algorithm is iterated until convergence (convergence is assessed by monitoring the change in the elements of  $\boldsymbol{\theta}^{(i)}$  for successive values of  $i$ ). For a comprehensive treatment of the EM algorithm, see [34].

A slight modification to the standard EM algorithm is necessary in this setting: Because  $\mathbf{Y}_f$  is lognormal, the E-step cannot be carried out analytically. However, since  $\zeta_t(\mathbf{s})$  is a Gaussian process, it turns out that we only need to compute the conditional expectation and the conditional covariance of  $(\mathbf{Y}_f, \mathbf{Y}_m) \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\theta}}$  (see (A.5) in Appendix A.1). We obtain these by finding the mode and the curvature of the log-density (thus invoking a Laplace approximation; see [35, p. 213]). Formulas required for the E-step are given in Appendix A.1. The resulting (approximate) algorithm is referred to as a Laplace-EM algorithm in [31].

For the models we consider, the optimisation (18) can be carried out using gradient descent. Since computing the derivatives of  $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$  can be quite costly, we suggest halting the gradient descent prematurely for the first few M-steps in order to decrease computational cost. The resulting ‘generalised EM algorithm’ [34, p. 84] converges to a stationary point  $\hat{\boldsymbol{\theta}}$  (when the E-step is exact), although it requires more iterations until convergence.

### 3.2. Computationally efficient models for the discrepancy term

As discussed in Section 2.1, one reason to employ a bivariate model is the flexibility it gives in choosing a model for the discrepancy term. Clearly, one should choose a model for the discrepancy that allows it to scale well with both  $|\mathcal{T}|$  and  $|D_m^L|$ . In this article we employ a separable space-time structure (e.g., [4, 36]) that is computationally efficient and is a reasonable choice in the face of lack of structural information on the discrepancy term  $\zeta_t(\mathbf{s})$ ; see Section 5 for further discussion of this choice.

A separable covariance function is one that can be characterised through

$$\text{cov}(\zeta_t(\mathbf{s}), \zeta_{t'}(\mathbf{u}) \mid \boldsymbol{\theta}) \equiv \sigma_\zeta^2 \rho_s(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\theta}) \rho_t(t, t' \mid \boldsymbol{\theta}),$$

where  $\rho_s(\cdot, \cdot \mid \boldsymbol{\theta})$  is a spatial correlation function and  $\rho_t(\cdot, \cdot \mid \boldsymbol{\theta})$  is a temporal correlation function. In this work, we restrict our attention to the following correlation functions:

$$\begin{aligned} \rho_s(\mathbf{s}, \mathbf{u} \mid d) &\equiv \exp(-\|\mathbf{u} - \mathbf{s}\|/d); \quad d > 0, \\ \rho_t(t, t' \mid a) &\equiv a^{|t' - t|}; \quad |a| < 1, \end{aligned}$$

where  $d$  and  $a$  are range-like parameters that determine the correlation length of the spatial and temporal components, respectively (we refer to  $d$  as a correlation e-folding length). The derivatives of these correlation functions with respect to  $d$  and  $a$  can be easily computed for the M-step. Recall that, in our model, the unknown parameters  $\boldsymbol{\theta}$  are all associated with the discrepancy term's covariance function; therefore,  $\boldsymbol{\theta} \equiv (\sigma_\zeta^2, a, d)'$ .

Separability implies that

$$\begin{aligned}\boldsymbol{\Sigma}_\zeta(\boldsymbol{\theta}) &\equiv (\text{cov}(\zeta_t(\mathbf{s}), \zeta_{t'}(\mathbf{u}) \mid \boldsymbol{\theta}) : \mathbf{s}, \mathbf{u} \in D_m^L; t, t' \in \mathcal{T}) \\ &= \sigma_\zeta^2 \mathbf{R}_{\zeta,t}(a) \otimes \mathbf{R}_{\zeta,s}(d),\end{aligned}$$

where  $\otimes$  is the Kronecker product,  $\mathbf{R}_{\zeta,t}(a)$  is a correlation matrix of size  $|\mathcal{T}| \times |\mathcal{T}|$  and  $\mathbf{R}_{\zeta,s}(d)$  is a correlation matrix of size  $|D_m^L| \times |D_m^L|$ . One can take advantage of the Kronecker product to greatly reduce the memory and computational requirements of the E-steps and M-steps. Formulas required for the M-step for this model are given in Appendix A.2.

### 3.3. Taking advantage of the closed-form gradient: Hamiltonian Monte Carlo

Recall that with an EHM, we first estimate the parameters using, for example, the EM algorithm. Then, after fixing them at the EM-estimate, we carry out inference on the bivariate fields using MCMC. This section describes the second, MCMC, stage for sampling from the empirical predictive distribution,  $[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ .

The most popular MCMC algorithm in use is a random-walk Metropolis algorithm, where one uses a symmetric density to propose new states in the Markov chain. Metropolis algorithms do not take into account the ‘shape’ of the density and, as a result, the optimal acceptance rate typically results in chains that are highly correlated. We can improve on the Metropolis sampler by using the gradient of the posterior density, which in our case can be computed analytically. The gradient is used to reduce the probability of proposing samples in the direction where the probability density drops off steeply, and to increase the probability in regions where the probability density is relatively smooth. Hamiltonian (or hybrid) Monte Carlo (HMC) is an MCMC method that takes into account gradient information and can thus be used to propose states that might be distant, in probability space, from the current sample and yet have a small chance of rejection. In practice, a trajectory for the samples is simulated using Hamiltonian dynamics [37].

A Hamiltonian system is described by a position vector and a momentum vector. In HMC,



the position vector corresponds to the current sample, while the momentum vector is randomly sampled from a pre-specified distribution. Hamilton’s equations are then used to simulate the trajectory of the current sample, under the random momentum vector, for a certain time interval. The potential-energy surface used is the negative log-density, and thus the trajectory of the sample can be thought of as a particle rolling inside a surface characterised by the probability density of interest. In practice, Hamilton’s equations are discretised in time using a suitable scheme like the leapfrog method [38], with a given discretisation interval  $\Delta$ . Then the number of time steps is  $L = (\text{time interval})/\Delta$ . Choice of  $\Delta$  and  $L$  are critical for good performance of the MCMC scheme.

In general,  $\Delta$  needs to be set such that the narrow, possibly curving valleys in the negative log-density can be explored, while  $L$  needs to be set such that the time interval,  $L\Delta$ , is large enough for the proposal to be treated as independent from the current sample. This is no simple task, however by transforming the variables to have roughly unit scale, we can explore values of  $L$  and  $\Delta$  for a suitable acceptance rate ( $\approx 60\%$ ) such that  $L\Delta \simeq 1$  [38]. The quantity  $\Delta$  can also be randomised in order to minimise the chance of obtaining periodicity in the trajectories; see [38] for further implementation details.

Since  $\mathbf{Y}_m$  is conditionally Gaussian, conditional on  $\mathbf{Y}_f, \hat{\boldsymbol{\theta}}$ , and  $\hat{\boldsymbol{\vartheta}}$ , we can marginalise out  $\mathbf{Y}_m$  from  $[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$  and use HMC to sample  $\{\mathbf{Y}_f^{(i)}\}$  directly from  $[\mathbf{Y}_f \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ . For each  $\mathbf{Y}_f^{(i)}$ , we then sample  $\mathbf{Y}_m^{(i)}$  from  $[\mathbf{Y}_m \mid \mathbf{Y}_f^{(i)}, \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ , which is Gaussian. This sampling scheme is known as a collapsed Gibbs sampler [39]. The log-density and the gradient for the collapsed distribution (i.e.,  $[\mathbf{Y}_f \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ ) can be derived in a very similar way to the Laplace-approximated E-step (Appendix A.1), and thus they are not detailed here.

## 4. Results

In this section, we implement the model of Section 2 and the inference techniques discussed in Section 3 in two settings. First, in Section 4.1 we use simulated datasets on a one-dimensional ‘toy’ problem. Second, in Section 4.2 we apply the method to inferring methane emissions from mole-fraction observations in the UK and Ireland, as described in Section 1.

### 4.1. Study with simulated data

In the simulation study we simulate the flux from some known prior distribution, and the interaction function, which describes the SRR, is synthesised from a spatio-temporal model.

For consistency with the analysis of methane emissions given in Section 4.2, we assume that the mole fractions obtained using the simulated flux field and the SRR are in ppb, that the spatial-length scales are in degrees, and that each time step corresponds to a temporal interval of 2 h. In order to facilitate the analysis and visualisation of the results, we carry out these simulations in one-dimensional space. Specifically, the domain of interest is  $D \equiv [-10^\circ, 10^\circ]$ , and  $D_f^L \equiv \{-9.9^\circ, -9.7^\circ, \dots, 9.9^\circ\}$ , so that  $|D_f^L| = 100$ . The lognormal spatial process that defines the flux field has  $\tilde{\mu}_f(s) \equiv 5$  and  $\tilde{C}_{ff}(s, u | \boldsymbol{\vartheta})$  given by a spherical, isotropic covariance function, with parameters  $\boldsymbol{\vartheta}$  set equal to those estimated in Section 2.3.

We synthesised an SRR that reflects something that would be obtained from a numerical model such as NAME. That is, we use an SRR characterised by a plume originating at  $u = s$  with a particular orientation that decays slowly with  $|u - s|$  (see Figure 1, top-right panel) and varies smoothly in time. We achieved this by simulating a flow process through a truncated normalised Gaussian density with spatio-temporally varying scale. Specifically, we first simulated a parameter  $v_t(s)$  from a Gaussian process with separable spatio-temporal covariance structure, and then we defined

$$b_t(s, u | v_t(s)) \equiv \exp\left(-\frac{(u - s)^2}{2v_t(s)^2}\right) I(|u - s| < |v_t(s)|) J(s, u), \quad (19)$$

where

$$J(s, u) \equiv \begin{cases} I[(u - s) \geq 0]; & v_t(s) \geq 0, \\ I[(u - s) \leq 0]; & v_t(s) < 0, \end{cases}$$

and  $I(\cdot)$  is the indicator function. In (19), the exponential function describes a bell-shaped curve centred at  $u = s$ , where the indicator function truncates this curve at  $u = s \pm v_t(s)$ . The third term,  $J(s, u)$ , then truncates the bottom half of the function if  $v_t(s) \geq 0$  and the upper half otherwise. A plot showing  $\{b_t(s, u) : u \in D_f^L; t \in \mathcal{T}\}$  at five locations  $s$  (that coincide with the observation locations,  $D_m^O$ , to be described later), where  $\mathcal{T} \equiv \{1, 2, \dots, T\}$ ,  $T = 100$  steps, is shown in Figure 3, top panel.

The SRR was subsequently used to construct the matrix  $\mathbf{B}_{B,t}$  defined by (6), with  $A(\mathbf{u}) = 0.2^\circ$ . For the purposes of simulation, we set the true-discrepancy parameters to  $\sigma_\zeta = 50$  ppb,  $a = 0.8$ , and  $d = 1^\circ$ , so that  $\boldsymbol{\theta} = (2500 \text{ ppb}^2, 0.8, 1^\circ)'$ . We also set  $E(\zeta_t(s)) = 0$ , thereby assuming that any background mole fraction has been correctly removed *a priori*. Using these parameters and assumptions, we then simulated the flux field from (8) and (9) and the mole-

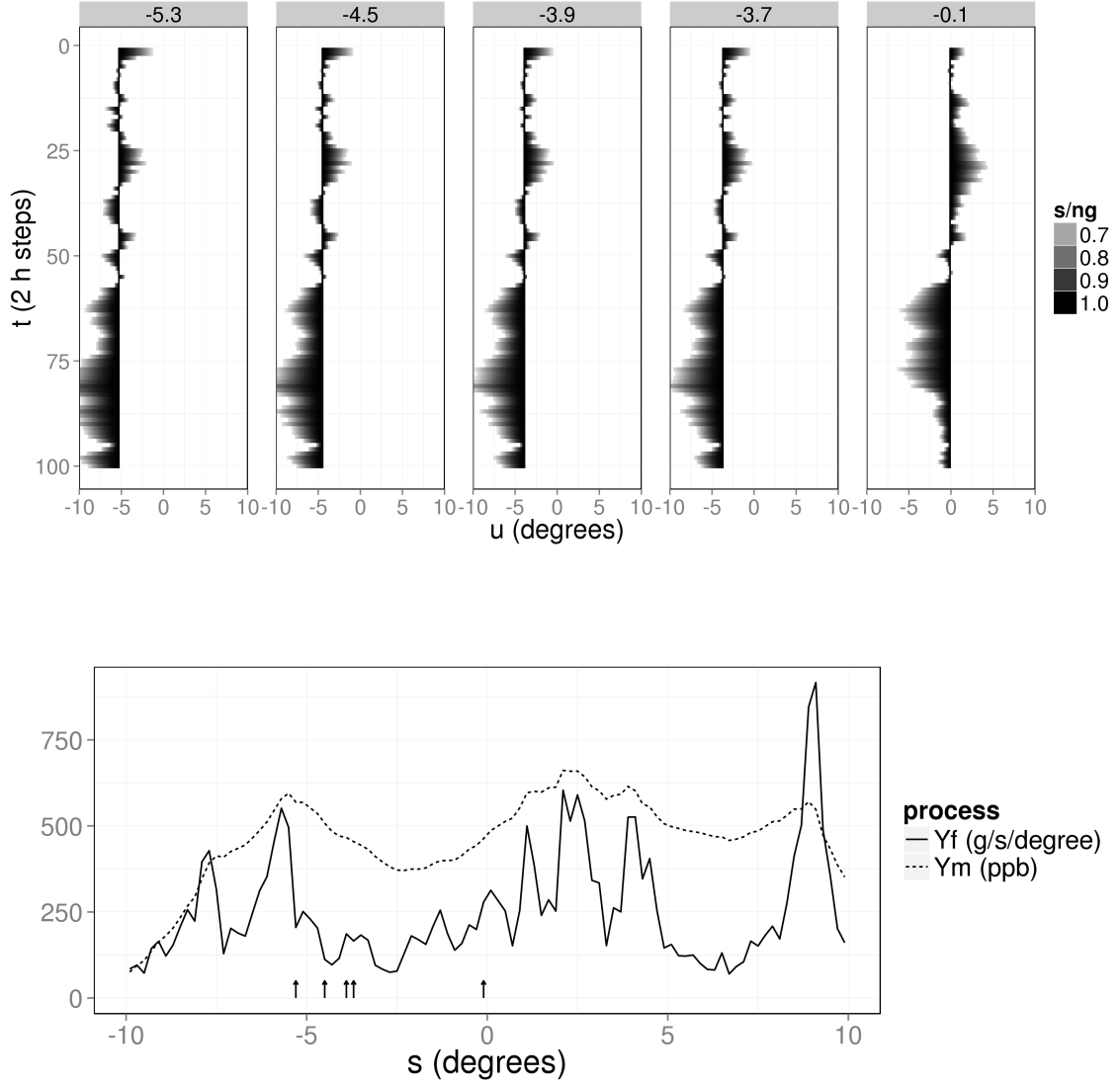


Figure 3: Top panel: The source-receptor relationship  $b_t(s, u)$  synthesised at five observation locations  $s \in D_m^O = \{-5.3^\circ, -4.5^\circ, -3.9^\circ, -3.7^\circ, -0.1^\circ\}$ . Note that  $b_t(s, u) = 0$  for  $u > 4.3$ . Bottom panel: A sample realisation of the flux field (solid line), the resulting time-averaged mole-fraction field (dashed line), and the five observation locations (arrows).

fraction field from (2). This simulation was done only once, to give us a simulated dataset where the true fields are known. In Figure 3, bottom panel, we show the true lognormal flux field and the true time-averaged mole-fraction field with  $D_m^L \equiv D_f^L$ .

We first considered the typical scenario of when we have only a few measurement stations where mole-fraction data are collected. For example, in Section 4.2, we have only four stations in the UK and Ireland. In this simulation study in one-dimensional space we considered five stations, shown by the arrows in the bottom panel of Figure 3, whose locations  $D_m^O$  were obtained by sampling randomly from  $D_f^L$ . Data were generated at these locations using (3),

and we assumed that data for all time steps were available (i.e.,  $D_{m,t}^O = D_m^O$ ;  $t \in \mathcal{T}$ ). Recall that  $T = 100$  time steps, which corresponds to about 8 days of 2-hourly observations. We assume that we are interested in predicting the mole-fraction field at the observation locations and, for illustration, at another location  $s = 0.3^\circ$ , and hence we set  $D_m^L = D_m^O \cup \{0.3^\circ\}$ .

The function  $b_t(s, u)$  is a measure of how sensitive  $Y_{m,t}(s)$  is to flux at  $u$ . It can be problematic, in ill-posed problems such as this one, to obtain inferences for locations  $u$  when the sensitivity at all measurement sites  $D_{m,t}^O$  is zero (or approximately zero) for  $t \in \mathcal{T}$ , since the influence of the flux field at these locations on the observed mole fractions is zero (or negligible). Without assuming a spatial process for the flux field, these regions would not even be identifiable, and hence they can be justifiably removed from the model. We therefore re-defined the flux domain  $D_f^L$  to exclude these regions. Formally, we can define  $D_t(s) \equiv \{u : b_t(s, u) > 0\}$  and re-define  $D_f^L \equiv \{-9.9^\circ, -9.7^\circ, \dots, 9.9^\circ\} \cap \left(\bigcup_{s \in D_m^O, t \in \mathcal{T}} D_t(s)\right)$ .

In order to run the Laplace-EM algorithm, we chose the starting value,  $\boldsymbol{\theta}^{(0)} \equiv (1000 \text{ ppb}^2, 0.2, 0.2^\circ)'$ , and commenced the gradient descent at  $\mathbf{Y}_f = \exp(\hat{\mu}_f + 0.5\hat{\sigma}_f^2)\mathbf{1}$  (where  $\hat{\mu}_f = 5$  and  $\hat{\sigma}_f^2 \equiv \tilde{C}_{ff}(s, s | \hat{\boldsymbol{\theta}})$ ) and  $\mathbf{Y}_m = \mathbf{Z}_m$ . Using 50 gradient descents at each M-step, the Laplace-EM algorithm converged in 40 iterations to  $\hat{\boldsymbol{\theta}} = (2000 \text{ ppb}^2, 0.71, 0.77^\circ)'$  (convergence was assessed visually). Recall that the true value of the parameters were  $\boldsymbol{\theta} = (2500 \text{ ppb}^2, 0.8, 1^\circ)'$ . We can expect the accuracy of these estimates to increase with the length of the observation record,  $T$ . Hence, when using this method in temporal blocks (e.g.,  $[4, 10]$ ), it is recommended to use large temporal intervals; for example in the case study in Section 4.2 we employ a three-month block with  $T \approx 1000$  time steps.

We substituted the EM-estimated  $\hat{\boldsymbol{\theta}}$  for  $\boldsymbol{\theta}$  into  $[\mathbf{Y}_f, \mathbf{Y}_m | \mathbf{Z}_m, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}]$  and sampled 10,000 times from the (empirical) posterior distribution of the flux field using the HMC sampler of Section 3.3 with  $L = 10$  steps and  $\Delta \in [0.066, 0.068]$ ; this resulted in an acceptance rate of 57%. Box plots showing inferences for the flux field from the HMC sampler (after discarding the first 1,000 samples for burn-in), together with the true flux field (crosses) and the observation locations (arrows on the horizontal axis) are shown in Figure 4. Notice how the uncertainty increases as distance away from the nearest observation location increases, as expected, but also notice how the true fluxes conform with our posterior inferences.

The Laplace approximation is used for parameter estimation and not for prediction of the fields, although predictions could be provided as by-products of the E-step at convergence of the EM algorithm. The reason we do not use the predictions from the Laplace approximation

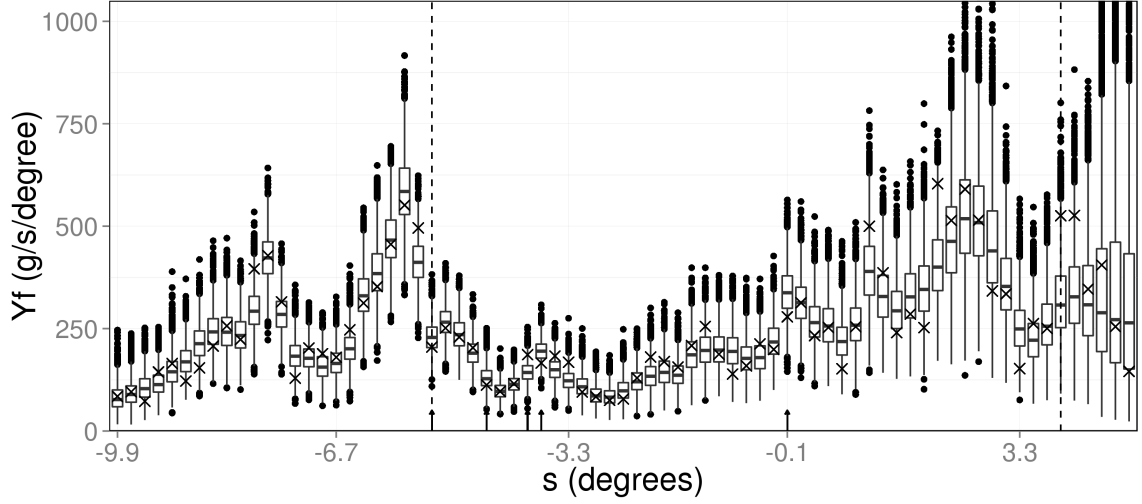


Figure 4: Samples from the posterior distribution of the lognormal flux field obtained using Hamiltonian Monte Carlo (HMC). The boxes denote the interquartile range, the whiskers extend to the last values that are within 1.5 times the interquartile range from the quartiles, and the dots show the samples that lie beyond the end of the whiskers. The crosses denote the true (simulated) fluxes, and the arrows along the horizontal axis denote the locations of the measurements. The vertical dashed lines show the spatial locations analysed in Figure 5. Since the observed mole fractions are insensitive to flux at  $\{s : s > 4.3\}$ , these locations were excluded from the model.

is that, while correctly locating the mode in the posterior distribution, these predictions do not account for skewness, something that is expected from a lognormal process. The difference between the prediction from the Laplace approximation and that from the HMC sampler at  $s = 3.9^\circ$  (far from an observation location) is shown in Figure 5, left panel. We find that this difference is less marked close to an observation location, since the posterior distribution is less sensitive to the prior distribution in these regions (e.g., at  $s = -5.3^\circ$ ); see Figure 5, right panel.

One of the contributions of this article is to introduce a lognormal-spatial model for the flux field, for trace gases where fluxes are positive. We show the importance of capturing the spatial characteristics of the flux field by comparing our results obtained with spatial range  $R = 3.3^\circ$ , to results we obtained with a misspecified, non-spatial model where  $R = 0.0^\circ$  in (16). To assess

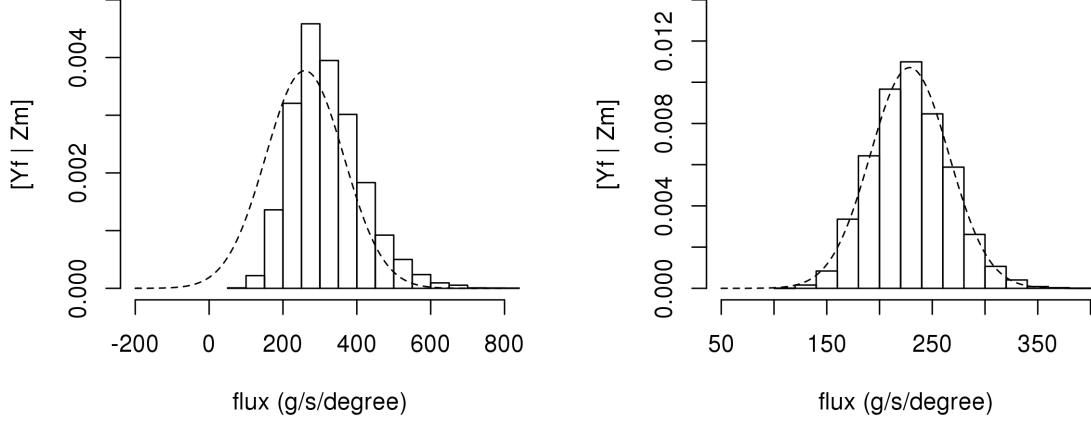


Figure 5: Laplace approximation (dashed line) and a histogram of the empirical posterior distribution from the MCMC samples for the flux at  $s = 3.9^\circ$  (left panel) and  $s = -5.3^\circ$  (right panel).

Table 1: Statistics  $S_{1,f}$ ,  $S_{2,f}$ , and  $S_{1,m}^{0.3}$  for the simulated case studies using a lognormal uncorrelated spatial process ( $R = 0.0^\circ$ ) and a lognormal correlated spatial process ( $R = 3.3^\circ$ )

$R$ (degrees)	$S_{1,f}$ ( $\text{g s}^{-1} \text{ degree}^{-1}$ )	$S_{2,f}$	$S_{1,m}^{0.3}$ (ppb)
0.0	97	0.74	32
3.3	62	0.83	32

the ‘goodness-of-prediction’ for both models we use the root-average-squared prediction error:

$$S_{1,f} = \left[ \frac{1}{|D_f^L|} \sum_{i=1}^{|D_f^L|} \left( Y_{f,i} - \hat{Y}_{f,i} \right)^2 \right]^{1/2}, \quad (20)$$

where  $Y_{f,i}$  is the  $i$ th element of the vector  $\mathbf{Y}_f$  and  $\hat{Y}_{f,i}$  is the posterior expectation of  $Y_{f,i}$  assuming a spatial range that may be different from the true range. The lower the value for  $S_{1,f}$ , the better the predictive performance in terms of pointwise prediction. To assess the distributional accuracy of the forecast, we use

$$S_{2,f} = \left[ \frac{\frac{1}{|D_f^L|} \sum_{i=1}^{|D_f^L|} \left( Y_{f,i} - \hat{Y}_{f,i} \right)^2}{\frac{1}{|D_f^L|} \sum_{i=1}^{|D_f^L|} \sigma_{Y_{f,i}}^2} \right]^{1/2}, \quad (21)$$

where  $\sigma_{Y_{f,i}}^2$  is the posterior variance of  $Y_{f,i}$ . Clearly,  $S_{2,f}$  should be close to 1 if uncertainty is correctly captured.

In Table 1, we show the statistics (20) and (21) obtained under uncorrelated (i.e., misspecified) and correlated (i.e., correctly specified) flux-field assumptions. The statistic  $S_{1,f}$  for the model with the spatially correlated flux was  $62 \text{ g s}^{-1} \text{ degree}^{-1}$ , while for the uncorrelated model (with  $R = 0.0^\circ$ ),  $S_{1,f}$  was  $97 \text{ g s}^{-1} \text{ degree}^{-1}$ . Assuming an uncorrelated flux field results in a considerable reduction in performance, caused mainly by poor observability of the flux field at  $s > 0.0^\circ$  (plot not shown). Because of such regions, it is important to capture the spatial correlations of the unobserved fields. Also, the statistic  $S_{2,f}$  is closer to 1 when  $R = 3.3^\circ$  than when  $R = 0.0^\circ$ . A value lower than 1 is indicative of under-confidence (an overall posterior variance which is too large); in our case,  $R = 0.0^\circ$  represents a model misspecification, and we see the consequence in the low value of  $S_{2,f}$  (namely, 0.74).

Another advantage of the proposed approach is the ease with which we can infer the mole-fraction field anywhere in the domain. In Figure 6, we show the distributions of the mole fractions at every time point at  $s = 0.3^\circ$ , which are consistent with the true mole fraction values. A statistic similar to (20) for evaluating the mole-fraction ‘goodness-of-prediction’ at  $s = 0.3^\circ$  is

$$S_{1,m}^{0.3} = \left[ \frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} \left( Y_{m,t}^{0.3} - \hat{Y}_{m,t}^{0.3} \right)^2 \right]^{1/2}, \quad (22)$$

where  $Y_{m,t}^{0.3}$  is the element of the vector  $\mathbf{Y}_{m,t}$  corresponding to the location  $s = 0.3^\circ$ , and  $\hat{Y}_{m,t}^{0.3}$  is the posterior expectation of  $Y_{m,t}^{0.3}$  assuming different values of  $R$ . Interestingly, unlike with  $S_{1,f}$ , the statistic (22) was found to be insensitive to  $R$ ; see Table 1. This reflects a well-known consequence of the ill-posed nature of the problem: There are several ‘plausible’ flux fields that can reproduce the observed mole fractions.

Next, we show that the bivariate approach scales well with dataset size by considering the case where  $|D_m^O| = 1000$ . The locations of the 1000 observations were obtained by uniform sampling within the domain, and the observations were generated using (3). Since the number of time steps is  $T = 100$ , we have 100,000 observations in all, far more than can be handled efficiently with the standard univariate approach based on (1). This problem becomes tractable by choosing  $D_m^L$  such that  $|D_m^L| \ll |D_m^O|$ , which is possible only when using a bivariate model. Here we set  $D_m^L = D_f^L$  and, since  $|D_f^L| = 100$ , we have 10,000 space-time locations at which to infer the mole-fraction field, which is an order of magnitude less than  $|D_m^O|$ . The Laplace-EM algorithm converged in 5 iterations to  $\hat{\theta} = (2300 \text{ ppb}^2, 0.81, 0.89)'$ . As expected, these estimates

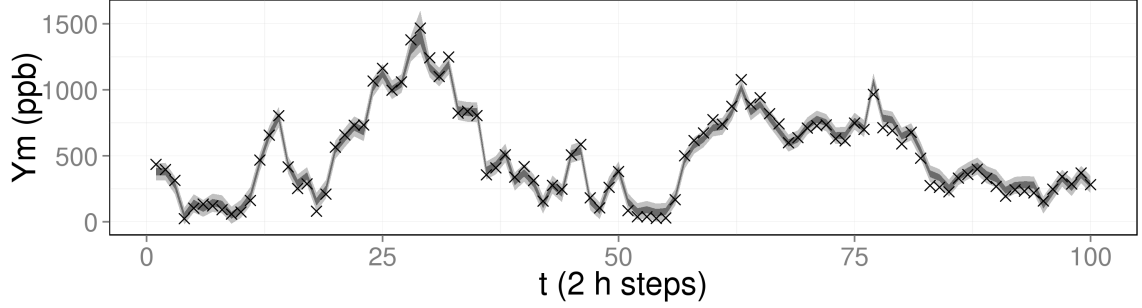


Figure 6: Distributions of mole fraction at  $s = 0.3^\circ$  and  $t \in \mathcal{T}$ , following Gibbs sampling. The dark and light shadings denote the interquartile and the 5th–95th percentile ranges, respectively. The crosses denote the true (simulated) mole fractions.

are more accurate due to the increased number of observations (recall that the true parameters were  $\theta = (2500 \text{ ppb}^2, 0.80, 1.00^\circ)'$ ). Increased accuracy was also noted in the posterior mode of the flux field, which was virtually identical to the true flux field at all locations in  $D_f^L$ .

Although it seems relatively straightforward to adopt this separable geostatistical model with large datasets, since both  $\mathbf{R}_{\zeta,t}$  and  $\mathbf{R}_{\zeta,s}$  are dense, one cannot use it when either  $|\mathcal{T}|$  or  $|D_m^L|$  become larger than, say, 2000. One way to remedy this limitation is to employ sparse inverse covariance matrices or reduced-rank covariance matrices instead; we provide further discussion on one such approach in Section 5.

Reproducible code for this simulation study is available from <https://github.com/andrewzm/atminv>.

#### 4.2. Inference on methane emissions in the UK and Ireland

In this section we present a case study that analyses real data on mole fraction in the UK and Ireland from the four stations shown in Figure 1, top-right panel, and we infer the discretised flux and mole-fraction fields.

##### 4.2.1. Data, preprocessing, and the source-receptor relationships

Methane observations, available in ppb, were made at four sites: Mace Head, Ireland (MHD), Ridge Hill, England (RGL), Tacolneston, England (TAC) and Angus, Scotland (TTA). These sites are from the UK Department of Energy and Climate Change (DECC) observation network [40]. In this analysis, the period January–March 2014 was used. The SRR was evaluated at the observation locations using the UK Met Office’s Numerical Atmospheric-dispersion Modelling Environment (NAME) [1], which is a Lagrangian Particle Dispersion Model (LPDM) that was



run backwards for 30 days. The spatial domain extended from  $-14^\circ$  to  $31^\circ$  E and  $36^\circ$  to  $66^\circ$  N at a grid resolution of  $0.352^\circ$  lon by  $0.234^\circ$  lat, thus covering the UK and a large part of continental Europe with a  $128 \times 128$  grid. For a complete description of the measurement procedure, ancillary information about the measurement sites, and a description of NAME and its application to prediction of UK and Ireland emissions, see [4].

Since the LPDM only simulates the effect of emissions from the previous 30 days before measurement, a substantial ‘background’ level of methane exists in the observed mole fractions that would not be accounted for by the LPDM. This background depends on the direction of the origin of the air mass due to, for example, the N-S latitudinal gradient or the vertical gradient. We used the following simple procedure to cater for the background. First, using NAME, we identified which observations were influenced by southerly transport or were influenced by the upper atmosphere due to vertical transport by tracking the directions that particles exited the NAME domain. Due to the complexities involved in estimating these backgrounds, these observations were removed from the dataset. Second, winds that were westerly in origin and passed over the Atlantic ocean were considered to be relatively free of regional emissions. Hence, this background, due to air masses originating in the west, can be clearly observed at Mace Head in Ireland. We used the 5th percentile of the extant mole fraction observations at Mace Head as a crude estimate of the background for all observations. Third, we subtracted this background estimate from the extant observations at all sites. In addition to removing data for background purposes, observations that could have significant influence from sub-grid scale processes (defined as ‘local’ processes), also identified using NAME, were removed. These data corresponded to times when there was a significant sensitivity to the nine grid cells surrounding the measurement point (i.e., when air was more likely to be stagnant).

To restrict our attention to land areas in the UK and Ireland, we used the methane flux inventory, together with NAME, to subtract contributions from the rest of Europe, including the sea territories of the UK and Ireland. Thus, we have final, extant observations of methane mole fraction, corrected for background and due to methane emissions only in the UK and Ireland land areas. The flux inventory and the NAME outputs were aggregated to a coarser resolution by grouping  $2 \times 2$  grid cells. This is a downsampling factor of four so that finally  $|D_f^L| = 122$  coarse-resolution grid cells over the UK and Ireland; see Figure 1, top panels. Since we were only provided with the interaction function  $b_t(\mathbf{s}, \cdot)$  evaluated with  $\mathbf{s}$  corresponding to the four measurement station locations, we set  $D_m^L = D_m^O$ .

#### 4.2.2. Inference

The final, extant dataset had 3,409 observations for the period January–March 2014. Out of these, we held 20% for validation purposes and used the other 80% for model-training purposes. The validation set was chosen by randomly sampling without replacement from all the observations.

In this application we are interested in computing the total flux by grid cell, in units  $\text{g s}^{-1}$ . This can be carried out directly by adding a grid-area-dependent offset to  $\tilde{\mu}_f(\mathbf{s} \mid \boldsymbol{\vartheta})$  and constructing  $\mathbf{B}_{B,t}$  without the grid-area weights (see Appendix B). The Laplace-EM algorithm was configured to carry out 10 gradient descents at each M-step, and it converged in 20 iterations to the parameter estimates,  $\hat{\boldsymbol{\theta}} \equiv (\hat{\sigma}_\zeta^2, \hat{a}, \hat{d})' = (690 \text{ ppb}^2, 0.972, 2.27^\circ)'$ . The standard deviation of the discrepancy,  $\hat{\sigma}_\zeta \simeq 26 \text{ ppb}$ , is considerably larger than that associated with fine-scale temporal variability during each 2 h window and instrumentation error ( $\simeq 8.6 \text{ ppb}$  at TAC, for example), and thus it is not negligible. Moreover, the magnitudes of  $\hat{d} = 2.27^\circ$  and  $\hat{a} = 0.972$  are indicative of considerable spatial and temporal dependence. The parameter  $\hat{a} = 0.972$  corresponds to a temporal-correlation e-folding length of  $-2/\ln(\hat{a}) = 70 \text{ h}$  (where the factor 2 adjusts for the temporal interval used, 2 h).

The case study presented here is illustrative, as it only considers a crude treatment of the background and makes the strong assumption that the fluxes outside the UK and Ireland land areas are known. Yet, it is indicative to compare our results to those in [4], where a similar model to ours was used, but without the lognormal spatial assumption. The spatial-correlation e-folding length obtained here ( $\hat{d} = 2.27^\circ$ ) matches within error that obtained in [4] (posterior median equal to 133 km), but the temporal-correlation e-folding length we obtain is 2–3 times larger than the one in [4]. Such a difference is possibly because they considered sea areas within the UK and Ireland territory (these were omitted from the model by us), because the result in [4] is averaged from several monthly studies over a 2-year period, and because we use a lognormal process for the flux field.

As in the simulation study (Section 4.1), we iterated the HMC sampler 10,000 times in order to obtain samples from the empirical predictive distribution,  $[\mathbf{Y}_f \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$ , with  $L = 20$  steps and  $\Delta \in [0.070, 0.071]$  in the HMC. These settings resulted in an acceptance rate of 54%. The pointwise posterior median and 95th percentile are shown in the left and right panel, respectively, of Figure 7, and they should be compared with the inventory emissions map shown in Figure 1, top-left panel. We obtain total estimates for the UK and Ireland (including inventory values

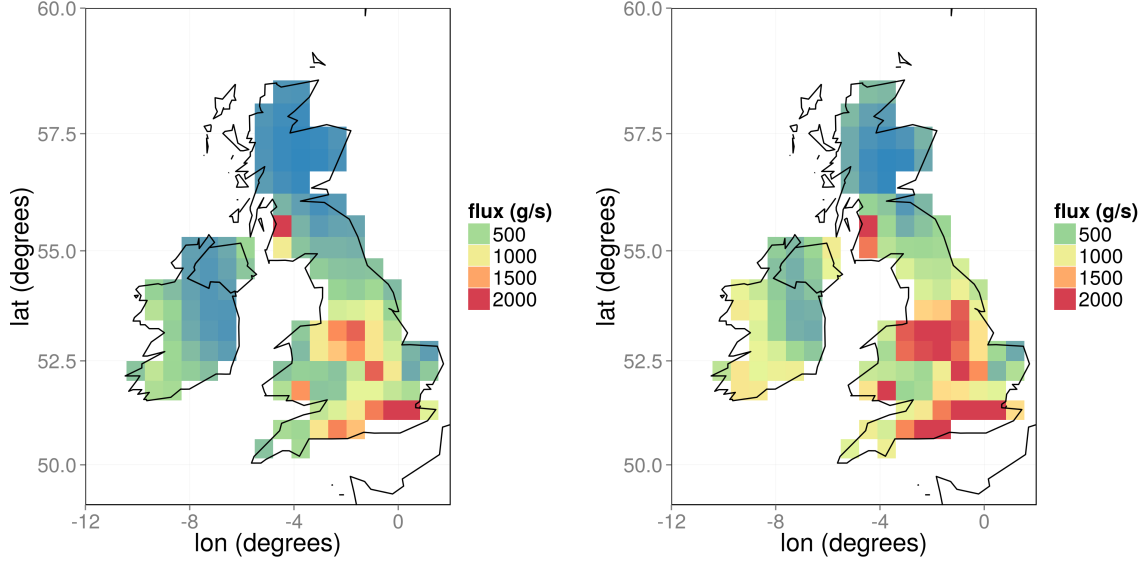


Figure 7: Median (left panel) and 95th percentile (right panel) total methane emissions (fluxes) in the UK and Ireland by grid cell, obtained using our Laplace-EM/HMC approach. White grid cells denote regions where total fluxes were assumed known and used to correct the observation as described in Section 4.2.1.

for the sea territories) of  $2.23 \pm 0.08 \text{ Tg yr}^{-1}$  and  $0.37 \pm 0.05 \text{ Tg yr}^{-1}$ , respectively. Both these estimates are lower than those obtained from the inventory shown in Figure 1, top-left panel ( $2.65 \text{ Tg yr}^{-1}$  and  $0.64 \text{ Tg yr}^{-1}$ , respectively), and they support the conclusion in [4] that these inventory emissions might be too large.

Interestingly, despite the omission of a spatially varying prior distribution, the posterior spatial distribution of emissions (Figure 7) is compatible somewhat with that in the inventory (Figure 1, top-left panel). However, there are some key differences. Many large cities, such as Edinburgh, Glasgow, and Dublin, were estimated with lower methane emissions. On the other hand, we obtained larger emissions for the south of the UK, most notably a median total emission of  $0.128 \text{ Tg yr}^{-1}$  in the grid cell containing London, which compares to  $0.062 \text{ Tg yr}^{-1}$  in the inventory. We note that there is no consensus on methane emissions in London, but our result supports the observation in [41] that there is a two- to three-fold difference between inventory and measured fluxes in central London.

In Figure 8, we show the distributions of mole fraction at each station location during the month of January, together with both those observations used to train the model (black crosses) and those observations left out for validation (red crosses). The first thing to note is that the distributions of mole fractions are available at times when observations are missing. Mole-fraction uncertainty increases at these times, but clear patterns are also discernible (e.g., at

RGL, the peak at  $t=130$  is noticeable). It would be straightforward to show the distribution of mole fractions at other isolated locations, however this requires the interaction function to be evaluated at these locations too. Second, sometimes the posterior density of the mole fraction at a space-time coordinate  $(\mathbf{s}, t)$  might have probability mass on the negative real axis. This does not imply a negative (unphysical) mole fraction, rather an imperfect correction of the background, which in this example is on the order of 1900 ppb. Third, the observations used for validation lie within the 90% prediction intervals, with the exception of a few outliers. The ability to obtain what seem to be realistic predictions for out-of-sample mole fractions increases our confidence in inferences for  $\mathbf{Y}_f$ . However, without the availability of validation methane-flux data, critical predictive performance measures in trace-gas inversion cannot be obtained.

## 5. Discussion

In this article, we present a bivariate spatio-temporal model for flux and mole-fraction fields that allows atmospheric trace-gas inversion. We make use of efficient computational methods, and we include spatial correlations in the flux-field description while maintaining lognormal properties typical of some trace gases. We give a way to estimate these spatial correlations using flux inventories, and we show how to include them within our model. Importantly, uncertainties are captured at each layer in the model (flux level, mole-fraction level, observation level), and predictive distributions of mole fractions include uncertainty on the SRR discrepancy.

However, as yet we do not capture the uncertainty in estimation of the parameters in the flux and discrepancy models. Consequently, adjustments might need to be made to the prediction intervals obtained from  $[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}]$  to cater for the uncertainty in  $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}\}$ . Of the several approaches that may be used for adjustment, the parametric bootstrap approach of Laird and Louis [42, 43, Chapter 3] appears to be readily applicable here. In the context of this work, one would forward-simulate an ensemble of mole-fraction observations using the parameter estimates  $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}\}$  and from each of these simulated observations, one would re-estimate the parameters using the Laplace-EM algorithm. Then the collection of parameter estimates can be used to obtain a collection of predictive densities (using HMC) that can be averaged to give an adjusted predictive distribution for  $\{\mathbf{Y}_f, \mathbf{Y}_m\}$ . Here we expect  $\boldsymbol{\vartheta}$  to be unidentifiable, so we suggest bootstrapping only to adjust for uncertainty in  $\boldsymbol{\theta}$ . Alternatively, one might obtain bootstrap estimates of  $\boldsymbol{\vartheta}$  using only the inventory data  $\mathbf{Z}_{inv}$  (Section 2.3).

The lognormal distribution was chosen in order to satisfy the positivity constraint typically

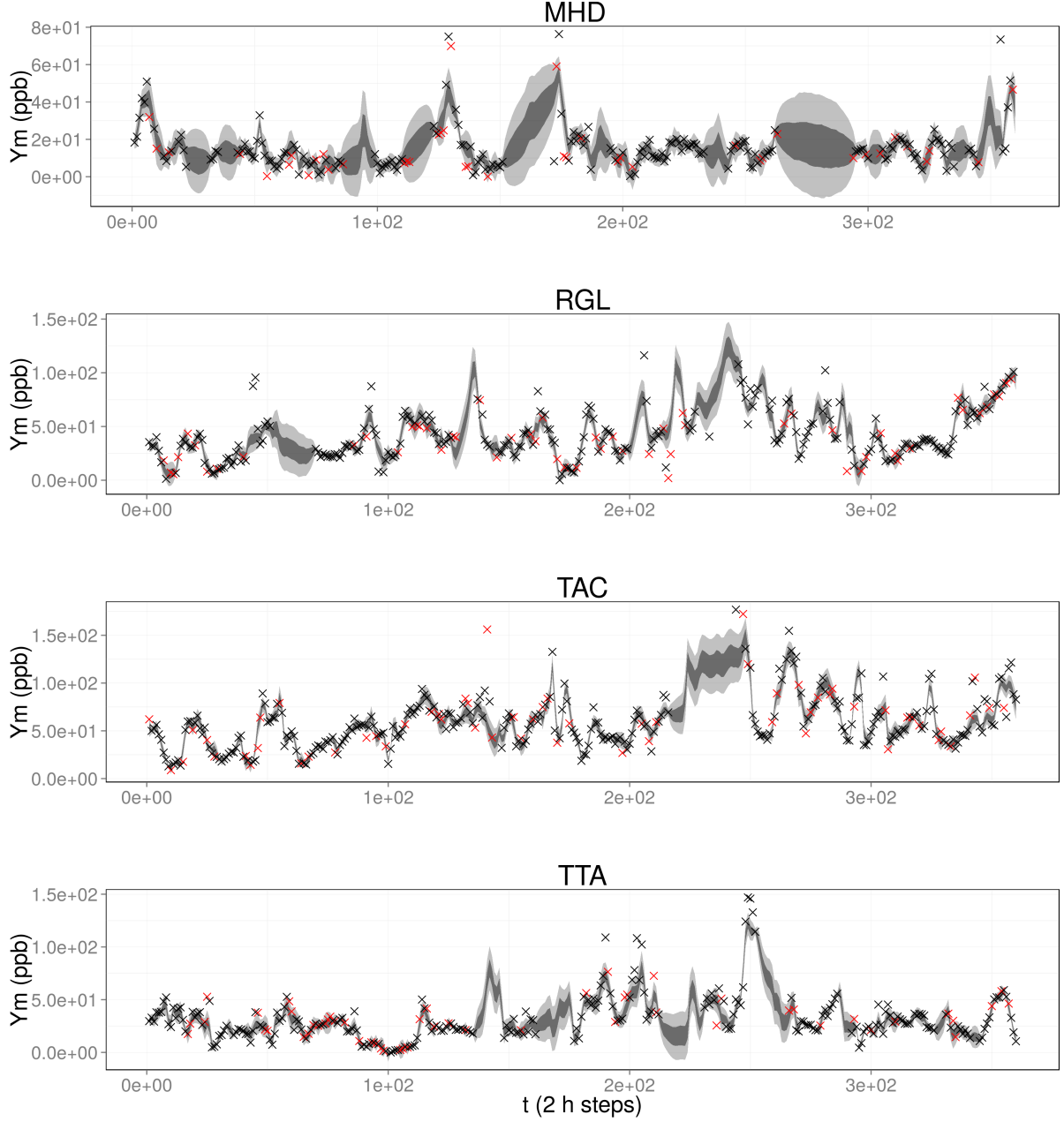


Figure 8: Distributions of mole fraction (with the background mole fraction subtracted) due to UK and Ireland land-based emissions at the four measurement stations, where  $t = 1$  corresponds to 01 January 2014 at 00:00,  $t = 2$  corresponds to 01 January 2014 at 02:00, and so on until 31 January 2014 at 22:00. The dark and light shadings denote the interquartile and the 5th–95th percentile ranges, respectively. The red crosses denote the observations used for validation, whilst the black crosses denote those used for training.

imposed on methane flux. As has been noted elsewhere [13], it has the undesirable property that negative fluxes occur with zero probability; negative methane fluxes are unlikely but can occur. This can be remedied by introducing a third, shifting parameter into the lognormal distribution such that  $P(Y_f \leq 0) > 0$ . The resulting distribution, known as the *three-parameter lognormal distribution* in the univariate case [44, p. 113], can also be calibrated using existing

inventories, in a manner similar to Section 2.3. This increases the model’s generality and, of course, distributions other than the lognormal may also be considered.

For computational reasons, we assumed that the discrepancy term has a covariance function that is separable in space and time. However, computational limits will also be reached with this approach, for increasing  $|\mathcal{T}|$  or  $|D_m^L|$ . The cost of storing an  $n \times n$  dense matrix is  $O(n^2)$ , and the computational cost of inverting it is  $O(n^3)$ . One approach to remedy this is to define sparse inverse covariance matrices for two Gaussian Markov random fields (GMRFs), one for space and one for time, over the space-time locations of the mole-fraction field; see [45] for a comprehensive treatment of GMRFs. This sparsity reduces considerably the memory and computational requirements in inference. Space-time separable GMRFs have been used in the context of air-quality monitoring in [46]. Another approach is through reduced-rank spatio-temporal models [47, 48, 49], and there spatio-temporal basis functions (but not separability) is required. Other possibilities are through predictive processes [50] and approximations to stochastic partial differential equations [51].

## Acknowledgements

Noel Cressie’s research was partially supported by the 2015–2017 Australian Research Council Discovery Project DP150104576 and partially supported by the National Aeronautics and Space Administration (NASA) grant NNH14ZDA001N-OCO2. Anita Ganesan was funded by the Natural Environment Research Council (NERC) Independent Research Fellowship NE/L010992/1. Measurements at Mace Head, Ridge Hill, Tacolneston, and Angus were funded by the UK Department of Energy and Climate Change (DECC) grant GA0201 to the University of Bristol. Measurements at Mace Head were also partially funded from NASA grant NNX11AF17G to the Massachusetts Institute of Technology, which supports the Advanced Global Atmospheric Gases Experiment (AGAGE) and at Tacolneston through the NERC National Centre for Atmospheric Research. The UK National Atmospheric Emissions Inventory (NAEI) was funded by DECC, the Department for Environment, Food and Rural Affairs (DEFRA), the Scottish Government, the Welsh Government, and the Northern Ireland Department of Environment. We would like to thank Matthew Rigby, Jonathan Rougier, and Dani Gamerman for discussions relevant to this article.

## References

- [1] A. Jones, D. J. Thomson, M. Hort, B. Devenish, The UK Met Office’s next-generation atmospheric dispersion model, NAME III, in: C. Borrego, A.-L. Norman (Eds.), *Air Pollution Modeling and its Application XVII*, Springer, New York, NY, 2007, pp. 580–589.
- [2] Department for Environment Food and Rural Affairs (DEFRA), National Atmospheric Emissions Inventory, <http://naei.defra.gov.uk>, 2014 [Online; accessed 01 March 2015].
- [3] Joint Research Centre of the European Commission (JRC) and the Netherlands Environmental Assessment Agency (PBL), Emission Database for Global Atmospheric Research (EDGAR), release version 4.2, <http://edgar.jrc.ec.europa.eu>, 2011 [Online; accessed 12 May 2014].
- [4] A. L. Ganesan, A. J. Manning, A. Grant, D. Young, D. E. Oram, W. T. Sturges, J. B. Moncrieff, S. O’Doherty, Quantifying methane and nitrous oxide emissions from the UK and Ireland using a national-scale monitoring network, *Atmospheric Chemistry and Physics* 15 (11) (2015) 6393–6406.
- [5] A. M. Michalak, L. Bruhwiler, P. P. Tans, A geostatistical approach to surface flux estimation of atmospheric trace gases, *Journal of Geophysical Research* 109 (2004) D14109.
- [6] J. Brynjarsdóttir, A. O’Hagan, Learning about physical parameters: the importance of model discrepancy, *Inverse Problems* 30 (2014) 114007.
- [7] M. C. Kennedy, A. O’Hagan, Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B* 63 (2001) 425–464.
- [8] M. G. Genton, W. Kleiber, Cross-covariance functions for multivariate geostatistics, *Statistical Science* 30 (2015) 147–163.
- [9] S. M. Miller, S. C. Wofsy, A. M. Michalak, E. A. Kort, A. E. Andrews, S. C. Biraud, E. J. Dlugokencky, J. Eluszkiewicz, M. L. Fischer, G. Janssens-Maenhout, B. R. Miller, J. B. Miller, S. A. Montzka, T. Nehrkorn, C. Sweeney, Anthropogenic emissions of methane in the United States, *Proceedings of the National Academy of Sciences* 110 (2013) 20018–20022.
- [10] A. L. Ganesan, M. Rigby, A. Zammit-Mangion, A. J. Manning, R. G. Prinn, P. J. Fraser, C. M. Harth, K.-R. Kim, P. B. Krummel, S. Li, J. Mühle, S. J. O’Doherty, S. Park, P. K.

- Salameh, L. P. Steele, R. F. Weiss, Characterization of uncertainties in atmospheric trace gas inversions using hierarchical Bayesian methods, *Atmospheric Chemistry and Physics* 14 (2014) 3855–3864.
- [11] N. Cressie, Block kriging for lognormal spatial processes, *Mathematical Geology* 38 (2006) 413–443.
- [12] N. Cressie, *Statistics for Spatial Data*, rev. edn, John Wiley & Sons, New York, NY, 1993.
- [13] S. M. Miller, A. M. Michalak, P. J. Levi, Atmospheric inverse modeling with known physical bounds: an example from trace gas emissions, *Geoscientific Model Development* 7 (2014) 303–315.
- [14] N. Cressie, C. K. Wikle, *Statistics for Spatio-Temporal Data*, John Wiley and Sons, Hoboken, NJ, 2011.
- [15] M. Rigby, A. J. Manning, R. G. Prinn, Inversion of long-lived trace gas emissions using combined Eulerian and Lagrangian chemical transport models, *Atmospheric Chemistry and Physics* 11 (2011) 9887–9898.
- [16] A. Stohl, P. Seibert, J. Arduini, S. Eckhardt, P. Fraser, B. R. Grealley, C. Lunder, M. Maione, J. Mühle, S. O’Doherty, R. G. Prinn, S. Reimann, T. Saito, N. Schmidbauer, P. G. Simmonds, M. K. Vollmer, R. F. Weiss, Y. Yokouchi, An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons, *Atmospheric Chemistry and Physics* 9 (2009) 1597–1620.
- [17] R. L. Thompson, A. Stohl, FLEXINVERT: an atmospheric Bayesian inversion framework for determining surface fluxes of trace species using an optimized grid, *Geoscientific Model Development* 7 (2014) 2223–2242.
- [18] C. G. Kaufman, M. J. Schervish, D. W. Nychka, Covariance tapering for likelihood-based estimation in large spatial data sets, *Journal of the American Statistical Association* 103 (2008) 1545–1555.
- [19] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, MA, 2006.
- [20] J. Aitchison, J. A. C. Brown, *The Lognormal Distribution with Special Reference to its Uses in Economics*, Cambridge University Press, Cambridge, UK, 1957.



- [21] K. R. Gurney, R. M. Law, A. S. Denning, P. J. Rayner, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, S. Fan, I. Y. Fung, M. Gloor, M. Heimann, K. Higuchi, J. John, T. Maki, S. Maksyutov, K. Masarie, P. Peylin, M. Prather, B. C. Pak, J. Randerson, J. Sarmiento, S. Taguchi, T. Takahashi, C.-W. Yuen, Towards robust regional estimates of CO<sub>2</sub> sources and sinks using atmospheric transport models, *Nature* 415 (2002) 626–630.
- [22] S. M. Gourdj, K. L. Mueller, K. Schaefer, A. M. Michalak, Global monthly averaged CO<sub>2</sub> fluxes recovered using a geostatistical inverse modeling approach: 2. Results including auxiliary environmental data, *Journal of Geophysical Research* 113 (2008) D21115.
- [23] N. Cressie, A. Zammit-Mangion, Bivariate spatial covariance models: A conditional approach, *submitted*, <http://arxiv.org/abs/1504.01865>, 2015 [Online; accessed 09 April 2015].
- [24] C. A. Calder, C. Berrett, T. Shi, N. Xiao, D. K. Munroe, Modeling space–time dynamics of aerosols using satellite data and atmospheric transport model output, *Journal of Agricultural, Biological, and Environmental Statistics* 16 (2011) 495–512.
- [25] A. Zammit-Mangion, J. Rougier, J. Bamber, N. Schön, Resolving the Antarctic contribution to sea-level rise: a hierarchical modelling framework, *Environmetrics* 25 (2014) 245–264.
- [26] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2014).
- [27] E. J. Pebesma, Multivariable geostatistics in S: the gstat package, *Computers & Geosciences* 30 (2004) 683–691.
- [28] C. K. Carter, R. Kohn, On Gibbs sampling for state space models, *Biometrika* 81 (1994) 541–553.
- [29] A. Zammit-Mangion, G. Sanguinetti, V. Kadirkamanathan, Variational estimation in spatiotemporal systems from continuous and point-process observations, *IEEE Transactions on Signal Processing* 60 (2012) 3449–3459.
- [30] E. L. Kang, D. Liu, N. Cressie, Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models, *Computational Statistics & Data Analysis* 53 (2009) 3016–3032.

- [31] A. Sengupta, N. Cressie, Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions, *Spatial Statistics* 4 (2013) 14–44.
- [32] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B* 39 (1977) 1–38.
- [33] M. Dewar, K. Scerri, V. Kadiramanathan, Data-driven spatio-temporal modeling using the integro-difference equation, *IEEE Transactions on Signal Processing* 57 (2009) 83–91.
- [34] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, NY, 1997.
- [35] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, 2006.
- [36] T. Gneiting, M. G. Genton, P. Guttorp, Geostatistical space-time models, stationarity, separability, and full symmetry, in: B. Finkenstadt, L. Held, V. Isham (Eds.), *Statistical Methods for Spatio-Temporal Systems*, Chapman & Hall/CRC Press, Boca Raton, FL, 2006, pp. 151–172.
- [37] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid Monte Carlo, *Physics Letters B* 195 (1987) 216–222.
- [38] R. Neal, MCMC using Hamiltonian dynamics, in: S. Brooks, A. Gelman, G. L. Jones, X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC Press, Boca Raton, FL, 2011, pp. 113–162.
- [39] D. A. Van Dyk, T. Park, Partially collapsed Gibbs samplers: theory and methods, *Journal of the American Statistical Association* 103 (2008) 790–796.
- [40] A. J. Manning, S. O’Doherty, Atmospheric trends, <http://www.metoffice.gov.uk/atmospheric-trends>, 2012 [Online; accessed 01 April 2015].
- [41] C. Helfter, E. Nemitz, J. F. Barlow, C. R. Wood, Carbon dioxide & methane emission dynamics in central London (UK), Poster presented at the EGU General Assembly, Vienna Austria. [http://www.picarro.com/sites/default/files/Helfter\\_EGU\\_2013\\_ERE51\\_urban\\_v3\\_0.pdf](http://www.picarro.com/sites/default/files/Helfter_EGU_2013_ERE51_urban_v3_0.pdf), 2013 [Online; accessed 01 March 2015].
- [42] N. M. Laird, T. A. Louis, Empirical Bayes confidence intervals based on bootstrap samples, *Journal of the American Statistical Association* 82 (1987) 739–750.

- [43] B. P. Carlin, T. A. Louis, Bayes and empirical Bayes methods for data analysis, 2nd Edition, Chapman & Hall/CRC Press, Boca Raton, FL, 2000.
- [44] N. L. Johnson, S. Kotz, Distributions in Statistics: Continuous Univariate Distributions–1, Houghton Mifflin, Boston, MA, 1970.
- [45] H. Rue, L. Held, Gaussian Markov Random Fields: Theory and Applications, Chapman and Hall/CRC Press, Boca Raton, FL, 2005.
- [46] M. Cameletti, F. Lindgren, D. Simpson, H. Rue, Spatio-temporal modeling of particulate matter concentration through the SPDE approach, *Advances in Statistical Analysis* 97 (2013) 109–131.
- [47] C. K. Wikle, R. F. Milliff, D. Nychka, L. M. Berliner, Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds, *Journal of the American Statistical Association* 96 (2001) 382–397.
- [48] N. Cressie, T. Shi, E. L. Kang, Fixed Rank Filtering for spatio-temporal data, *Journal of Computational and Graphical Statistics* 19 (2010) 724–745.
- [49] M. Katzfuss, N. Cressie, Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets, *Journal of Time Series Analysis* 32 (2011) 430–446.
- [50] S. Banerjee, A. E. Gelfand, A. O. Finley, H. Sang, Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society: Series B* 70 (2008) 825–848.
- [51] F. Lindgren, H. Rue, J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society B* 73 (2011) 423–498.

## **Appendix A. The Laplace-EM algorithm for estimating discrepancy parameters**

We first describe some of the notation used throughout this appendix. The symbol  $\oslash$  will be used to denote element-wise division whilst the symbol  $\odot$  denotes element-wise multiplication. These operators can only be used for vectors or matrices of the same size and result in a vector, or matrix, of that size. Throughout, we will use numerator layout notation: The derivative of a scalar by a (column) vector returns a transposed vector, whilst the derivative of a scalar by

a matrix  $\mathbf{X}$  returns a matrix whose  $(i, j)$ th element corresponds to the derivative of the scalar with respect to the  $(j, i)$ th element of  $\mathbf{X}$ .

We shall also be repeatedly using the following identities for two vectors  $\mathbf{X}, \mathbf{Y}$  in the E-step:

$$\begin{aligned}\text{diag}(\mathbf{Y})\mathbf{X} &\equiv \text{diag}(\mathbf{X})\mathbf{Y}, \\ \frac{\partial \ln \mathbf{X}}{\mathbf{X}} &\equiv \text{diag}(\mathbf{1} \oslash \mathbf{X}),\end{aligned}$$

where  $\text{diag}(\mathbf{X})$  returns a diagonal matrix with  $\mathbf{X}$  on the diagonal.

In the M-step of the EM algorithm we shall use the following identities for two matrices  $\mathbf{A}, \mathbf{B}$  of size  $p \times p$  and  $q \times q$  respectively, and scalar  $x$ :

$$\begin{aligned}|\mathbf{A}^n| &\equiv |\mathbf{A}|^n, \\ (\mathbf{A} \otimes \mathbf{B})^{-1} &\equiv \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}, \\ |\mathbf{A} \otimes \mathbf{B}| &\equiv |\mathbf{A}|^q |\mathbf{B}|^p, \\ \frac{\partial \ln |\mathbf{A}|}{\partial x} &\equiv \text{tr} \left( \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \right), \\ \frac{\partial (\mathbf{A} \otimes \mathbf{B})}{\partial x} &\equiv \mathbf{A} \otimes \frac{\partial \mathbf{B}}{\partial x} + \frac{\partial \mathbf{A}}{\partial x} \otimes \mathbf{B},\end{aligned}\tag{A.1}$$

$$\tag{A.2}$$

where in this context  $|\cdot|$  is the determinant and  $\otimes$  is the Kronecker product.

#### Appendix A.1. E-Step

In the E-step, we compute

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) \equiv \mathbb{E}(\ln L_c(\boldsymbol{\theta}) \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}),\tag{A.3}$$

that is, the expectation of the complete-data log-likelihood with respect to the conditional distribution of  $\{\mathbf{Y}_f, \mathbf{Y}_m\}$  given the data  $\mathbf{Z}_m, \hat{\boldsymbol{\vartheta}}$ , and the current parameter estimate  $\boldsymbol{\theta}^{(i)}$ ,  $i = 0, 1, \dots$ . Hence, in order to be able to evaluate  $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$ , we first need to determine the conditional distribution under which to take expectations.

Define  $\mathbf{Q}_O \equiv \sigma_\varepsilon^{-2} \mathbf{I}$ ,  $\mathbf{Q}_\zeta(\boldsymbol{\theta}^{(i)}) \equiv \boldsymbol{\Sigma}_\zeta^{-1}(\boldsymbol{\theta}^{(i)})$ , and  $\tilde{\mathbf{Q}}_f(\hat{\boldsymbol{\vartheta}}) = \tilde{\boldsymbol{\Sigma}}_f^{-1}(\hat{\boldsymbol{\vartheta}})$ , where

$$\begin{aligned}\boldsymbol{\Sigma}_\zeta(\boldsymbol{\theta}^{(i)}) &\equiv (\text{cov}(\boldsymbol{\zeta}_t(\mathbf{s}), \boldsymbol{\zeta}_{t'}(\mathbf{u}) \mid \boldsymbol{\theta}^{(i)}) : \mathbf{s}, \mathbf{u} \in D_m^L; t, t' \in \mathcal{T}), \\ \tilde{\boldsymbol{\Sigma}}_f(\hat{\boldsymbol{\vartheta}}) &\equiv (\tilde{C}_{ff}(\mathbf{s}, \mathbf{u} \mid \hat{\boldsymbol{\vartheta}}) : \mathbf{s}, \mathbf{u} \in D_f^L).\end{aligned}$$

For notational convenience, from now on we shall omit the dependence of  $\tilde{\Sigma}_f$  and  $\tilde{\mathbf{Q}}_f$  on  $\hat{\boldsymbol{\vartheta}}$ , and the dependence of  $\Sigma_\zeta$  and  $\mathbf{Q}_\zeta$  on  $\boldsymbol{\theta}^{(i)}$ .

The observation model for the discretised system is,

$$\mathbf{Z}_m = \mathbf{C}\mathbf{Y}_m + \boldsymbol{\varepsilon}_m,$$

where  $\mathbf{C} \equiv (\mathbf{C}'_t : t \in \mathcal{T})'$  is the incidence matrix mapping the discretised mole-fraction field to the observations. Define  $\mathbf{B}_B \equiv (\mathbf{B}'_{B,t} : t \in \mathcal{T})'$  and  $\tilde{\boldsymbol{\mu}}_f \equiv (\tilde{\mu}_f(\mathbf{s} \mid \hat{\boldsymbol{\vartheta}}) : \mathbf{s} \in D_f^L)'$ . Then

$$\begin{aligned} \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}] &= c_1 - 0.5(\mathbf{Z}_m - \mathbf{C}\mathbf{Y}_m)' \mathbf{Q}_O (\mathbf{Z}_m - \mathbf{C}\mathbf{Y}_m) \\ &\quad - 0.5(\mathbf{Y}_m - \mathbf{B}_B \mathbf{Y}_f)' \mathbf{Q}_\zeta (\mathbf{Y}_m - \mathbf{B}_B \mathbf{Y}_f) \\ &\quad - 0.5(\ln \mathbf{Y}_f - \tilde{\boldsymbol{\mu}}_f)' \tilde{\mathbf{Q}}_f (\ln \mathbf{Y}_f - \tilde{\boldsymbol{\mu}}_f) - (\ln \mathbf{Y}_f)' \mathbf{1}. \end{aligned} \tag{A.4}$$

Due to the presence of  $\ln \mathbf{Y}_f$  in (A.4), this conditional distribution is not Gaussian. We summarise it using the first two moments, approximated using first-order and second-order derivatives.

Define  $\mathbf{D}_f \equiv \text{diag}(\mathbf{1} \oslash \mathbf{Y}_f)$  and  $\mathbf{D}_{ff} \equiv \text{diag}(\mathbf{1} \oslash (\mathbf{Y}_f \odot \mathbf{Y}_f))$ . Then the derivatives  $\mathbf{J}$  of (A.4) with respect to  $\mathbf{Y} \equiv (\mathbf{Y}'_f, \mathbf{Y}'_m)'$ , are given by

$$\begin{aligned} \mathbf{J}(\mathbf{Y} \mid \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}) &\equiv \begin{pmatrix} \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}'_f} \\ \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}'_m} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}'_B \mathbf{Q}_\zeta \mathbf{Y}_m - \mathbf{B}'_B \mathbf{Q}_\zeta \mathbf{B} \mathbf{Y}_f - \mathbf{D}_f \tilde{\mathbf{Q}}_f (\ln \mathbf{Y}_f - \tilde{\boldsymbol{\mu}}_f) - (\mathbf{1} \oslash \mathbf{Y}_f) \\ -\mathbf{C}' \mathbf{Q}_O \mathbf{C} \mathbf{Y}_m + \mathbf{C}' \mathbf{Q}_O \mathbf{Z}_m - \mathbf{Q}_\zeta (\mathbf{Y}_m - \mathbf{B}_B \mathbf{Y}_f) \end{pmatrix}. \end{aligned}$$

The Hessian  $\mathbf{H}$  is

$$\begin{aligned} \mathbf{H}(\mathbf{Y} \mid \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}) &\equiv \begin{pmatrix} \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}_f \mathbf{Y}_f'} & \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}_m \mathbf{Y}_f'} \\ \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}_f \mathbf{Y}_m'} & \frac{\partial \ln[\mathbf{Y}_f, \mathbf{Y}_m \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}]}{\partial \mathbf{Y}_m \mathbf{Y}_m'} \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{B}_B' \mathbf{Q}_\zeta \mathbf{B}_B - \mathbf{D}_f \tilde{\mathbf{Q}}_f \mathbf{D}_f + \text{diag}(\tilde{\mathbf{Q}}_f (\ln \mathbf{Y}_f - \tilde{\boldsymbol{\mu}}_f)) \mathbf{D}_{ff} + \mathbf{D}_{ff} & \mathbf{B}_B' \mathbf{Q}_\zeta \\ \mathbf{Q}_\zeta \mathbf{B}_B & -\mathbf{C}' \mathbf{Q}_O \mathbf{C} - \mathbf{Q}_\zeta \end{pmatrix}. \end{aligned}$$

We now use a gradient-descent algorithm with the derivatives  $\mathbf{J}(\mathbf{Y} \mid \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}})$  to find the mode  $\mathbf{Y}^*$  of the distribution and use this as an approximation to the mean. We evaluate  $\mathbf{H}$  at this mode; the approximate covariance is then given by  $-\mathbf{H}(\mathbf{Y}^* \mid \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}})^{-1}$ . For gradient descent, we used the  $R$  function `optim` with the method `BFGS`.

From (A.3), the function  $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$  is

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) = -0.5 \ln |\boldsymbol{\Sigma}_\zeta(\boldsymbol{\theta})| - 0.5 \text{tr}(\mathbf{Q}_\zeta(\boldsymbol{\theta}) \boldsymbol{\Psi}),$$

where

$$\begin{aligned} \boldsymbol{\Psi} &\equiv \mathbb{E}(\mathbf{Y}_m \mathbf{Y}_m' \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}) \\ &\quad + \mathbf{B}_B \mathbb{E}(\mathbf{Y}_f \mathbf{Y}_f' \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}) \mathbf{B}_B' \\ &\quad - 2 \mathbf{B}_B \mathbb{E}(\mathbf{Y}_f \mathbf{Y}_m' \mid \mathbf{Z}_m, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\vartheta}}). \end{aligned} \tag{A.5}$$

Note that only the first two moments of  $(\mathbf{Y}_f', \mathbf{Y}_m')'$  are required to compute (A.5).

#### Appendix A.2. M-step

In the M-step we set

$$\boldsymbol{\theta}^{(i+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}). \tag{A.6}$$

The discrepancy term  $\zeta_t(\mathbf{s})$  has a separable spatio-temporal covariance function. Hence, we can write the full space-time covariance matrix as

$$\begin{aligned} \boldsymbol{\Sigma}_\zeta(\boldsymbol{\theta}) &= \sigma_\zeta^2 \mathbf{R}_\zeta(a, d) \\ &= \sigma_\zeta^2 \mathbf{R}_{\zeta,t}(a) \otimes \mathbf{R}_{\zeta,s}(d), \end{aligned} \tag{A.7}$$

where recall that  $\otimes$  is the Kronecker product and  $\mathbf{R}_{\zeta,s}$ ,  $\mathbf{R}_{\zeta,t}$  are correlation matrices. In the M-step we maximise  $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$  in (A.6) using gradient descent; the following gradients are obtained by repeated application of (A.1) and (A.2):

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})}{\partial \sigma_{\zeta}^2} \equiv 0.5 \left( -n\sigma_{\zeta}^{-2} + \sigma_{\zeta}^{-4} \text{tr}(\mathbf{R}_{\zeta}(a, d)^{-1} \boldsymbol{\Psi}) \right),$$

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})}{\partial a} &\equiv -0.5 \text{tr} \left( (\mathbf{R}_{\zeta,t}(a) \otimes \mathbf{R}_{\zeta,s}(d))^{-1} \left( \frac{\partial \mathbf{R}_{\zeta,t}(a)}{\partial a} \otimes \mathbf{R}_{\zeta,s}(d) \right) \right) \\ &\quad + 0.5 \sigma_{\zeta}^{-2} \text{tr} \left( \mathbf{R}_{\zeta}^{-1}(a, d) \left( \frac{\partial \mathbf{R}_{\zeta,t}(a)}{\partial a} \otimes \mathbf{R}_{\zeta,s}(d) \right) \mathbf{R}_{\zeta}^{-1}(a, d) \boldsymbol{\Psi} \right), \end{aligned}$$

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})}{\partial d} &\equiv -0.5 \text{tr} \left( (\mathbf{R}_{\zeta,t}(a) \otimes \mathbf{R}_{\zeta,s}(d))^{-1} \left( \mathbf{R}_{\zeta,t}(a) \otimes \frac{\partial \mathbf{R}_{\zeta,s}(d)}{\partial d} \right) \right) \\ &\quad + 0.5 \sigma_{\zeta}^{-2} \text{tr} \left( \mathbf{R}_{\zeta}^{-1}(a, d) \left( \mathbf{R}_{\zeta,t}(a) \otimes \frac{\partial \mathbf{R}_{\zeta,s}(d)}{\partial d} \right) \mathbf{R}_{\zeta}^{-1}(a, d) \boldsymbol{\Psi} \right), \end{aligned}$$

where here  $n = |D_m^L| |\mathcal{T}|$ . In the equations above, the mixed-product property of the Kronecker product can be used to simplify the computations; for example,

$$(\mathbf{R}_{\zeta,t}(a) \otimes \mathbf{R}_{\zeta,s}(d))^{-1} \left( \frac{\partial \mathbf{R}_{\zeta,t}(a)}{\partial a} \otimes \mathbf{R}_{\zeta,s}(d) \right) \equiv \left( \mathbf{R}_{\zeta,t}^{-1}(a) \frac{\partial \mathbf{R}_{\zeta,t}(a)}{\partial a} \right) \otimes \mathbf{I}.$$

All that remains is to specify the partial derivatives of the spatial and temporal correlation matrices with respect to  $a$  and  $d$ . These are obtained from the derivatives of the correlation functions:

$$\begin{aligned} \frac{\partial}{\partial a} \rho_t(t, t') &\equiv |t - t'| a^{|t-t'|-1}; \quad |a| < 1, \\ \frac{\partial}{\partial d} \rho_s(\mathbf{s}, \mathbf{u}) &\equiv \frac{\|\mathbf{u} - \mathbf{s}\|}{d^2} \exp\left(-\frac{\|\mathbf{u} - \mathbf{s}\|}{d}\right); \quad d > 0. \end{aligned}$$

## Appendix B. Directly inferring the total flux by grid cell

Instead of computing the flux (a density that is per unit area), running the EM-Laplace/HMC algorithm, and rescaling again to obtain the total flux, we can instead infer the total flux by grid cell,  $\{A(\mathbf{u})Y_f(\mathbf{u}) : \mathbf{u} \in D_f^L\}$ , directly with only a few modifications. Write down the

approximation to the integral in (2) as

$$\begin{aligned} \int_D b_t(\mathbf{s}, \mathbf{u}) Y_f(\mathbf{u}) d\mathbf{u} &\simeq \sum_{\mathbf{u} \in D_f^L} b_t(\mathbf{s}, \mathbf{u}) (A(\mathbf{u}) Y_f(\mathbf{u})) \\ &= \sum_{\mathbf{u} \in D_f^L} b_t(\mathbf{s}, \mathbf{u}) Y_f^{tot}(\mathbf{u}), \end{aligned}$$

where  $Y_f^{tot}(\mathbf{u}) \equiv A(\mathbf{u}) Y_f(\mathbf{u})$ . Now we seek a model for  $Y_f^{tot}(\mathbf{u})$ ; however, this poses no additional work since

$$\tilde{Y}_f^{tot}(\mathbf{s}) \equiv \ln Y_f^{tot}(\mathbf{s}) = \ln A(\mathbf{s}) + \ln Y_f(\mathbf{s}). \quad (\text{B.1})$$

It thus follows that  $Y_f^{tot}$  is also a lognormal spatial process, shifted by a known quantity. From (B.1),

$$\begin{aligned} \mathbb{E}(\tilde{Y}_f^{tot}(\mathbf{s}) \mid \boldsymbol{\vartheta}) &= \ln A(\mathbf{s}) + \tilde{\mu}_f(\mathbf{s} \mid \boldsymbol{\vartheta}), \\ \text{cov}(\tilde{Y}_f^{tot}(\mathbf{s}), \tilde{Y}_f^{tot}(\mathbf{u}) \mid \boldsymbol{\vartheta}) &= \tilde{C}_{ff}(\mathbf{s}, \mathbf{u} \mid \boldsymbol{\vartheta}). \end{aligned}$$

Note that the covariance function remains unchanged, and in order to solve directly for  $\tilde{Y}_f^{tot}(\cdot)$  one uses  $\tilde{\mu}_f^{tot}(\mathbf{s}) \equiv \ln A(\mathbf{s}) + \tilde{\mu}_f(\mathbf{s} \mid \boldsymbol{\vartheta})$ , instead of  $\tilde{\mu}_f(\mathbf{s} \mid \boldsymbol{\vartheta})$  in (8) and (9). Then  $\mathbf{B}_{B,t}$  is constructed *without* the area-weight terms since they have already been accounted for. That is, in this case the matrix  $\mathbf{B}_{B,t}$  is given by,

$$\mathbf{B}_{B,t} \equiv (b_t(\mathbf{s}, \mathbf{u}) : \mathbf{s} \in D_m^L; \mathbf{u} \in D_f^L).$$